Journal of Educational Psychology

RAYMOND G. KUHLEN, EDITOR
Syracuse University

Advisory Editors

JOHN B. CARROLL
Harvard University

LEE J. CRONBACH
University of Illinois

FRANCIS J. DIVESTA
Syracuse University

N. L. GAGE
Stanford University

CHESTER W. HARRIS
University of Wisconsin

ERNEST R. HILGARD
Stanford University

DOROTHEA MCCARTHY
Fordham University

CARSON MCGUIRE
University of Texas

PAUL H. MUSSEN
University of California, Berkeley

T. ERNEST NEWLAND
University of Illinois

H. H. REMMERS
Purdue University

JULIAN C. STANLEY
University of Wisconsin

J. B. STROUD
State University of Iowa

DONALD E. SUPER
Teachers College, Columbia University

GEORGE G. THOMPSON
Ohio State University

ROBERT L. THORNDIKE
Teachers College, Columbia University

BENTON J. UNDERWOOD
Northwestern University

VOLUME 57, 1966

HELEN ORR
Managing Editor

HENRY S. SALMON
Advertising and Business Manager

ELIZABETH MULLEN
Technical Editor

ROBERT D. GORMAN
Subscription Manager

PUBLISHED BIMONTHLY BY:
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 Seventeenth Street, N. W.
WASHINGTON, D. C. 20036

$10.50

Copyright, 1965, by the American Psychological Association, Inc.

Bureau of Educl. & Psyl. Research
D. C. E. R. T.
6 3.81
Date

CONTENTS OF VOLUME 57

Alexakos, C. E. Predictive Efficiency of Two Multivariate Statistical Techniques in Comparison with Clinical Predictions 297
Ali, Fahmida. See McCandless, Boyd R.
Alter, Richard D. See Harvey, O. J.
Amble, Bruce R., and Muehl, Siegmar. Perceptual Span Training and Reading Achievement of School Children 192
Amster, Harriett. Effect of Instructional Set and Variety of Instances on Children's Learning 74
Anderson, Harry E., Jr., White, William F., and Wash, James A. Generalized Effects of Praise and Reproof 169
Berlyne, D. E. Conditions of Prequestioning and Retention of Meaningful Material 128
Boersma, Frederick J. Effects of Delay of Information Feedback and Length of Postfeedback Interval on Linear Programed Learning 140
Carrier, Neil A., and Jewell, Donald O. Efficiency in Measuring the Effect of Anxiety upon Academic Performance 23
Cattell, Raymond B. See Horn, John L.
Coats, William D., and Smidchens, Uldis. Audience Recall as a Function of Speaker Dynamism 189
Crist, Robert L. Overt versus Covert Responding and Retention by Sixth-Grade Students 99
Dizney, Henry F. See Yamamoto, Kaoru
Duncanson, James P. Learning and Measured Abilities 220
Frase, Lawrence T. Validity Judgments of Syllogisms in Relation to Two Sets of Terms 239
Freyberg, P. S. Concept Development in Piagetian Terms in Relation to School Attainment 164
Friesen, Wallace V. See Kounin, Jacob S.
Goodwin, William L. Effect of Selected Methodological Conditions on Dependent Measures Taken after Classroom Experimentation 350
Grinder, Robert E. Relations of Social Dating Attractions to Academic Orientations and Peer Relations 27
Harari, Herbert, and McDavid, John W. Cultural Influences on Retention of Logical and Symbolic Material 18
Hart, J. T. Methodological Note on Feeling-of-Knowing Experiments 347
Hartnett, Rodney T. See Lehmann, Irvin J.
Harvey, O. J., White, B. Jack, Prather, Misha S., Alter, Richard D., and Hoffmeister, James K. Teachers' Belief Systems and Preschool Atmospheres 373
Haywood, H. Carl, and Wachs, Theodore D. Size-Discrimination Learning as a Function of Motivation-Hygiene Orientation in Adolescents 279
Hoffmeister, James K. See Harvey, O. J.
Holland, James G. See Kemp, Frederick D.
Horn, John L., and Cattell, Raymond B. Refinement and Test of the Theory of Fluid and Crystallized General Intelligences 253
Hudgins, Bryce B., and Smith, Louis M. Group Structure and Productivity in Problem-Solving 287
Iscoe, Ira. See Semler, Ira J.
Jeffrey, W. E. See Samuels, S. J.
Jewell, Donald O. See Carrier, Neil A.
Johnson, Donald M., and Stratton, R. Paul. Evaluation of Five Methods of Teaching Concepts 48
Kagan, Jerome, Pearson, Leslie, and Welch, Lois. Modifiability of an Impulsive Tempo 359
Kelly, Richard. Comparison of the Effects of Positive and Negative Vicarious Reinforcement in an Operant Learning Task 307

Kemp, Frederick D., and Holland, James G. Blackout Ratio and Overt Responses in Programed Instruction: Resolution of Disparate Results 109

Kight, Howard R., and Sassenrath, Julius M. Relation of Achievement Motivation and Test Anxiety to Performance in Programed Instruction 14

Kounin, Jacob S., Friesen, Wallace V., and Norton, A. Evangeline. Managing *Emotionally Disturbed Children in Regular Classrooms* 1

Kremenak, Shirley. See Muehl, Siegmar.

Lawton, Marcia J. Delay of Gratification as a Function of Characteristics of Social Agents 246

Lehmann, Irvin J., Sinha, Birendra K., and Hartnett, Rodney T. Changes in Attitudes and Values Associated with College Attendance 89

Lott, Albert J., and Lott, Bernice E. Group Cohesiveness and Individual Learning. 61

Lott, Bernice E. See Lott, Albert J.

Lynch, Steve. See Rohwer, William D., Jr.

Mace, Lawrence. Sequence of Vocal Response-Differentiation Training and Auditory Stimulus-Discrimination Training in Beginning French 102

McCandless, Boyd R., and Ali, Fahmida. Relations among Physical Skills and Personal and Social Variables in Three Cultures of Adolescent Girls 366

McDavid, John W. See Harari, Herbert.

Meyers, C. E. See Orpet, R. E.

Muehl, Siegmar. See Amble, Bruce R.

Muehl, Siegmar, and Kremenak, Shirley. Ability to Match Information within and between Auditory and Visual Sense Modalities and Subsequent Reading Achievement 230

Norton, A. Evangeline. See Kounin, Jacob S.

Olson, David R., and Pau, A. S. Emotionally Loaded Words and the Acquisition of a Sight Vocabulary 174

Orpet, R. E., and Meyers, C. E. Six Structure-of-Intellect Hypotheses in Six-Year-Old Children 341

Overing, Robert L. R., and Travers, Robert M. W. Effect upon Transfer of Variations in Training Conditions 179

Owens, W. A. Age and Mental Abilities: A Second Adult Follow-Up 311

Pau, A. S. See Olson, David R.

Pearson, Leslie. See Kagan, Jerome.

Prather, Misha S. See Harvey, O. J.

Rand, Leonard P. See Richards, James M., Jr.

Rand, Lorraine M. See Richards, James M., Jr.

Reynolds, James H. Cognitive Transfer in Verbal Learning 382

Richards, James M., Jr., Rand, Lorraine M., and Rand, Leonard P. Description of Junior Colleges 207

Rohwer, William D., Jr., and Lynch, Steve. Semantic Constraint in Paired-Associate Learning 271

Rotberg, Iris C. See Seidel, Robert J.

Samuels, S. J., and Jeffrey, W. E. Discriminability of Words, and Letter Cues Used in Learning to Read 337

Samuels, S. Jay. Effect of Experimentally Learned Word Associations on the Acquisition of Reading Responses 159

Sassenrath, Julius M. See Kight, Howard R.

Schoer, Lowell. Reactive Inhibition as Related to Performance on Programed Learning Materials 86

Seidel, Robert J., and Rotberg, Iris C. Effects of Written Verbalization and Timing of Information on Problem Solving in Programed Learning 151

Semler, Ira J., and Iscoe, Ira. Structure of Intelligence in Negro and White Children. 326

Sinha, Birendra K. See Lehmann, Irvin J.

Smidchens, Uldis. See Coats, William D.

Smith, Louis M. See Hudgins, Bryce B.

Stratton, R. Paul. See Johnson, Donald M.

Thistlethwaite, Donald L., and Wheeler, Norman. Effects of Teacher and Peer Subcultures upon Student Aspirations 35
Thorndike, Robert L. Intellectual Status and Intellectual Growth 121
Traub, Ross E. Importance of Problem Heterogeneity to Programed Instruction 54
Travers, Robert M. W. See Overing, Robert L. R.
Wachs, Theodore D. See Haywood, H. Carl
Wash, James A. See Anderson, Harry E., Jr.
Welch, Lois. See Kagan, Jerome.
Wheeler, Norman. See Thistlethwaite, Donald L.
White, B. Jack. See Harvey, O. J.
White, William F. See Anderson, Harry E., Jr.
Williams, Joanna P. Combining Response Modes in Programmed Instruction 215
Yamamoto, Kaoru, and Dizney, Henry F. Eight Professors—A Study on College Students' Preferences among their Teachers 146
Yonge, George D. Structure of Experience and Functional Fixedness 115
Zaslow, Robert W. Reversals in Children as a Function of Midline Body Orientation. 133

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 57 February 1966 Number 1

CONTENTS

Managing Emotionally Disturbed Children in Regular Classrooms 1
JACOB S. KOUNIN, WALLACE V. FRIESEN, AND A. EVANGELINE NORTON

Relation of Achievement Motivation and Test Anxiety to Performance in Programed Instruction 14
HOWARD R. KIGHT AND JULIUS M. SASSENRATH

Cultural Influences on Retention of Logical and Symbolic Material 18
HERBERT HARARI AND JOHN W. MCDAVID

Efficiency in Measuring the Effect of Anxiety upon Academic Performance 23
NEIL A. CARRIER AND DONALD O. JEWELL

Relations of Social Dating Attractions to Academic Orientations and Peer Relations 27
ROBERT E. GRINDER

Effects of Teacher and Peer Subcultures upon Student Aspirations 35
DONALD L. THISTLETHWAITE AND NORMAN WHEELER

Evaluation of Five Methods of Teaching Concepts 48
DONALD M. JOHNSON AND R. PAUL STRATTON

Importance of Problem Heterogeneity to Programed Instruction 54
ROSS E. TRAUB

Journal of Educational Psychology
1966, Vol. 57, No. 1, 1–13

MANAGING EMOTIONALLY DISTURBED CHILDREN IN REGULAR CLASSROOMS[1]

JACOB S. KOUNIN
Wayne State University

WALLACE V. FRIESEN
San Francisco State College

AND A. EVANGELINE NORTON
Wayne State University

This study was an attempt to delineate some dimensions of teaching style that affect the behavior of emotionally disturbed children in Grades 1–2 and 3–5 in regular classrooms. Video tapes were obtained for ½ day of 30 classrooms, each containing 1 or more emotionally disturbed children. Scores for deviancy and work involvement were obtained for both disturbed (E) and nondisturbed (Non E) children. Some findings are: (a) Pupils' scores vary between seatwork and recitation settings in the same classroom; (b) E children manifest less school-appropriate behavior than Non E children; (c) teachers who are successful in managing the behavior of Non E children are also relatively successful with E children; (d) techniques of handling misbehavior as such do not correlate with children's behavior; (e) teacher "With-it-ness," techniques of handling group movement, and programing for variety change in learning activities do correlate with children's behavior.

Although authorities recognize that emotionally disturbed children are present in schools, and that the classroom is important in the world of the child, we know of no definitive research on what the classroom teacher can *do* in programing for, or in managing the behavior of, emotionally disturbed children in regular classrooms. Much of the literature in this area concerns the teacher's selecting children to refer for treatment by some other person, agency, or special class; her providing information to these agents; her introducing some sort of unit on "human relations" or "family living"; or her attending a course, seminar, or workshop in human relations or mental hygiene. Other writings recommend selecting teachers who have certain kinds of personalities, characteristics, or traits such as patience, understanding, flexibility, sensitivity, insight, etc. No doubt the advocated characteristics are desirable, but little has been written about how these translate into techniques a teacher can use in managing a classroom.

The present research is an attempt to discover techniques that teachers can utilize to manage the overt behavior of emotionally disturbed children in regular classrooms. A further concern is whether these techniques are significant in the management of emotionally disturbed children only, or whether they are also significant in the management of normal children.

After a year spent observing regular classrooms containing emotionally disturbed children, the authors were impressed by (*a*) the differences in the behavior of these children with different teachers, (*b*) a tendency to "label" both teachers and children, (*c*) a tendency to note perceptually outstanding aspects of a classroom and to attribute the behavior of children to these, and (*d*) the general complexity and variety of events occurring in a classroom. The authors concluded that this type of research required a medium that would provide a nonselective, complete, and objective record of all observable events in a classroom, one that would allow many viewings of the same events so that a variety of variables—for the teacher and subsettings as well as for the children—could be studied. The method chosen should not only provide a means of measuring the validity of existing hypotheses and notions, but should also provide data on variables or hypotheses unformulated at the time of data gathering.

---

[1] This investigation was supported by Public Health Service Research Grant MH-04221 from the National Institute of Mental Health.

DATA-GATHERING PROCEDURE

The following method was used for gathering data:

1. Two TV cameras were mounted in nearly soundproof boxes and placed on stands in a classroom. One camera, kept stationary, was placed in a position to record the largest possible area of the classroom continuously and without "editing"; the other camera, equipped with a pan-tilt mechanism, enabled the engineeer to record various positions in the classroom. The decision about what to record and where to superimpose this picture upon the major picture was made by the project director. Priority was given to video taping the teacher or the emotionally disturbed children when they were "lost" from the main picture.
2. All recording was controlled by engineers using equipment in a truck parked outside the school building.
3. Two microphones were used to record sound; one was a wireless radio transmitting microphone worn by the teacher that made possible the recording of everything she said.

Video tapes of classrooms were obtained from two predominantly middle-class suburbs. Suburb X was involved in a study aimed at evaluating special classes, in the first two grades, for emotionally disturbed children. Referred children were individually administered approximately 14 hours of psychological, educational, psychiatric, and neurological tests. Some children were assigned to special classes (experimental children), while "matched" children (control cases) remained in regular classes. Video tapes were obtained for (*a*) the two special classes containing 14 experimental children, but not included in this report, and (*b*) the 15 regular classrooms containing the 14 matched children along with 24 children diagnosed as emotionally disturbed by means of the same diagnostic procedures, but not assigned to experimental or control groups at the time this study was conducted.

In Suburb Y, 15 classrooms in Grades 3–5 containing one or more emotionally disturbed children (totalling 32) were video taped. The children were selected from a list, provided by principals, of children who had been given individual psychological and psychiatric tests by the staff of a nearby children's psychiatric hospital.

All classrooms were video taped continuously, for an entire morning or afternoon, except when the children left the room, as for recess or gymnasium periods.

CODING PROCEDURES

Only behavior in academic activities—reading, arithmetic, social studies, science, etc.—was coded. Child behavior during nonacademic activities (attendance, milk time, nonacademic games) was not coded. Scores were kept separately for four types of academic subsettings: (*a*) subgroup recitation periods (Sub Rec), where the teacher worked with a subgroup of the class (usually a reading group) while the remainder of the children worked independently at their seats; (*b*) entire-class recitation (En Rec) periods, where the teacher worked with the entire class in a recitation period (explaining, demonstrating, questioning, etc.); (*c*) seatwork nonsphere periods (SW non Sp), where a group of children engaged in some kind of seatwork without direct supervision by the teacher at the time (occurred in conjunction with Sub Rec); and (*d*) seatwork sphere (SW Sp), where the entire class engaged in seatwork while the teacher worked with them (circulating, supervising, occasionally commenting, etc.).

The authors' question about the children's behavior was simple: To what degree is their behavior school appropriate? The authors attempted to measure child behavior which would reflect teachers' success in (*a*) getting the children to engage in the classroom activities (in this case, the academic learning activities), and (*b*) keeping them from misbehaving and disrupting the legitimate behavior of others.

This paper reports on two of these child-behavior codes: work involvement and deviancy rate.

The behavior of from two to eight nondisturbed children (Non E), in addition to that of each emotionally disturbed child (E), was coded for each of the four subsettings. The Non E children were selected to represent the four quadrants of the seating arrangement during the activity. The number of Non E children whose behavior was coded varied according to the number of children in the subsetting (some Sub Rec groups contained as few as two children), and the number that were clearly visible. All E children were coded throughout the academic subsettings in which they were present and their behavior was codable.

The behavior being coded was categorized every 10 seconds. The categories were: (*a*) definitely and completely involved in work (DI), (*b*) probably involved in work (PI), (*c*) definitely not involved in work (DO), (*d*) restless, (*e*) languishing, (*f*) languishing and restless, (*g*) engaged in task-related deviancy (TR), (*h*) engaged in non-task-related deviancy (NTR) that was mild or serious. In the case of the E children, the contagion of their deviancy was coded: Whenever an E child engaged in a NTR, the behavior of the children in this vicinity was also coded at the time.

Scores for each classroom and each subsetting were kept separately for E and Non E children. Three scores will be discussed here: work-involvement rate, or the mean DI/DO ratio; deviancy rate, or the mean percentage of 10-second intervals that contain a tally for NTR—mild NTR being scored 1, serious NTR being scored 2 (Fewer than 1% of the intervals in the regular classrooms contained a serious deviancy—overt aggressive acts, open defiance of teacher—and these were found in but five of the classrooms.); and contagion rate of NTR's for E children, or the mean percentage of children in the vicinity of the E who entered into a deviancy initiated by the E.

Intercoder agreement between various pairs of the four coders ranged from 82%–96% agreement with a mean of 92%.

The teacher-style codes selected for consideration in this paper will be described in the Results section.

Insofar as possible, the authors attempted to adhere to the following principles in developing codes for teacher style: (*a*) The categories should designate concrete behaviors, that is, specific managerial technologies, rather than "traits" or "characteristics" of teachers; (*b*) dimensions should be those on which teachers differed. For example, although all teachers designate times for children to make a transition from one activity to another, not all teachers do it in the same manner; (*c*) the dimensions should be as free as possible of commitment to any philosophical orientation or value system.

RESULTS

Table 1 presents a summary of the work-involvement and deviancy scores for both the E and Non E children in the four major activity subsettings. (The *N*s vary between subsettings because not all teachers programed all four subsettings. The *N*s vary between E and Non E groups because E children were not always present in the

TABLE 1

Mean Work-Involvement and Deviancy Scores of Emotionally Disturbed and Nondisturbed Children in Four Classroom Subsettings

Subsetting	Emotionally disturbed children		Nondisturbed children		*t*
	Score	*N*	Score	*N*	
Ratio					
Definitely in/definitely not involved in work					
Subgroup recitation	76.4	19	88.7	21	2.66*
Class recitation	59.6	19	78.4	19	2.67*
Seatwork in teacher's sphere	50.6	14	67.5	14	2.18*
Seatwork not in teacher's sphere	44.8	20	60.5	21	2.28*
Non-task-related deviancy score					
Subgroup recitation	2.6	19	1.9	21	5.49**
Class recitation	5.9	19	2.8	19	2.95**
Seatwork in teacher's sphere	15.4	14	5.6	14	2.26*
Seatwork not in teacher's sphere	13.0	20	10.3	21	.79

* $p < .05$.
** $p < .01$.

subsetting coded or because an adequately visible sample of Non E children was not always present.)

In all subsettings, the behavior of the Non E children was more school appropriate than that of the E children. This difference apeared in both work-involvement scores and deviancy scores.

The behavior of *all* children varied with the classroom setting. The highest degree of school-appropriate behavior occurred in subgroup recitation periods (mostly reading groups), and the lowest degree of such behavior occurred in seatwork situations. The differences between combined recitation periods and seatwork periods were significant at the .01 level or better for both the DI/DO ratio and the NTR score and for both E and Non E children.

It can be concluded that, although the school behavior of E and Non E children differs in degree of school appropriateness, both types of children change similarly in behavior according to the classroom situation.

Are special techniques necessary to manage the behavior of emotionally disturbed children in regular classrooms—techniques that are different from those used to manage other children? Or are the necessary managerial techniques essentially the same for all children in a classroom? A partial answer to this question can be obtained by seeing whether E and Non E children respond in the same manner to different teachers.

The correlations between the behavior of E and Non E children are presented in Table 2. Because of subsetting differences, these are presented for the four subsettings separately and for recitation periods combined and seatwork periods combined. Due to differences between Grades 1–2 and Grades 3–5 (i.e., NTR rate tended to be lower and DI/DO ratios higher in the upper grades), results are given for all grades combined and for the lower grades separately.

In view of the r's shown in Table 2, it can be concluded that the degree of school-appropriate behavior manifested by Non E children is positively related to the degree of school-appropriate behavior exhibited by E children in the same classroom. Insofar as the behavior of the children can be attributed to the managerial success of the teacher, teachers who are successful in managing a classroom are also relatively successful in managing the emotionally disturbed children present in that classroom.

TABLE 2

SPEARMAN r'S BETWEEN THE BEHAVIOR OF EMOTIONALLY DISTURBED AND NONDISTURBED CHILDREN

Subsetting	All grades				Grades 1 & 2			
	Deviancy rate	*N*	Work-involvement ratio	*N*	Deviancy rate	*N*	Work-involvement ratio	*N*
Seatwork in teacher's sphere	.381	14	.398	14	—	3	—	3
Seatwork not in teacher's sphere	.641**	18	.581**	18	.806**	10	.488	10
All seatwork	.477**	25	.611**	25	.697*	10	.600*	10
Subgroup recitation	.558**	18	.194	17	.906**	10	.750**	9
Class recitation	.701**	19	.554*	17	—	6	—	6
All recitation	.777**	27	.638**	25	.856**	13	.734**	12

* $p < .05$; one-tailed test.

** $p < .01$; one-tailed test.

Teachers sometimes express concern that emotionally disturbed children disrupt the behavior of others. Is this disruptive influence a consequence of the deviant behavior of emotionally disturbed children or is it generally symptomatic of unsuccessful classroom management?

In an attempt to answer this question, we estimated as follows the contagion effect of the misbehavior of E children: Whenever an E child initiated deviant behavior, the children in his vicinity were coded as "getting in" with it or as remaining with school-appropriate behavior. The percentage of children disrupted by the E child's misbehavior was determined, and an average contagion rate per incident was computed for each classroom. This was compiled for all SW settings combined and correlated with the deviancy rate of Non E children in all SW settings. (Rec settings were not used here because of the low frequency of E child deviancies in Rec subsettings.) For all grades combined, the correlation was .829 ($N = 25$; $p < .01$). For Grades 1 and 2 the correlation was .718 ($N = 10$; $p < .01$).

It may be concluded that the degree of contagion of an E child's deviant behavior is related to the degree of misbehavior occurring in the classroom. The teacher who is generally successful in managing the surface behavior of children in her classroom not only has a lower deviancy rate among emotionally disturbed children in that classroom, but also produces a classroom climate that "contains" the misbehavior of an E child and prevents it from disrupting the behavior of other children.

Are there concrete teaching techniques from which one can predict the school-appropriate behavior of children? If so, do these techniques have the same effect upon the behavior of both emotionally disturbed and non-disturbed children?

The first, or Desist, code dealt with what one might call "direct" managerial techniques: the techniques of dealing with misbehavior as such. Every time the teacher did something to stop a child's misbehavior, the incident was coded for how the teacher did it and the degree of success she had in stopping the misbehavior.

The degree of success was coded on a 7-point scale. This ranged from the target child's continuing to misbehave on one end, to the child's stopping the misbehavior immediately *and* engaging in the required work on the other. Intercoder agreement averaged 74%; agreement within 1 point averaged 95%.

The Desist technique was rated on (*a*) clarity, (*b*) firmness ("I mean it" quality), (*c*) child treatment (negative, neutral, positive), (*d*) intensity, and (*e*) focus (misbehavior, i.e., "stop that talking," or task, i.e., "do those arithmetic problems," or both misbehavior and task). Intercoder agreement averaged 90%.

For only 1 of 30 teachers was there a significant correlation between any aspect of her Desist technique and the degree of immediate success of the effort. Thus, for any one teacher, the degree of clarity, firmness, etc., of her behavior during the event did not relate to the target child's stopping his deviancy or taking up the prescribed task.

Teachers were also ranked according to their mean Desist-style scores for clarity, firmness, punitiveness, etc. These rankings did not correlate with the rankings on deviancy or work involvement for either their E or their Non E children.

It may be concluded that the manner of dealing with misbehavior as such is not a significant determinant of how well children behave in a classroom. This generalization should be limited to the type of misbehavior here considered (i.e., fewer than 1% of deviancies were rated as "serious"), the kinds of Desist techniques used by the teachers (i.e., while high anger did occur, no extreme punitiveness was observed, such as whipping or expelling during the times classrooms were video taped), and the length of time the classrooms had been in existence (i.e., video tapes were taken in approximately the eighth month of school). A history of firmness, clarity, or punitiveness in desisting may have operated to reduce the significance of these variables. Kounin and Gump (1958) found that clarity, firmness, and roughness did make a difference in conformity-deviancy of children who were observers of Desist events during the first 4 days of kindergarten; they also found more "no observable effects" during the second through fourth day than on the first day.

Desist events were coded for two other categories of teacher style which could be judged from such events, namely, "Overlappingness" and "With-it-ness."

Overlappingness refers to the degree to which a teacher overtly manifests signs of attention to two aspects of a classroom simultaneously. When two issues are confronted, does she attend to both or drop one completely and go "all out" to the other? For example, she may be working with a reading group when a deviancy occurs in the seatwork group. She may issue a Desist order to the child in the seatwork group while simultaneously attending to the reading group (remaining in the same seat, instructing the reader to continue reading, scanning the reader while desisting the deviant). Or, she may immerse herself in the deviant and completely drop the reading group for the time (changing manner and geographically moving away without any instructions to, or scanning of, the reading group).

With-it-ness refers to the degree to which the teacher *communicates* to the class the impression that she does or does not know what is going on (has the proverbial "eyes in back of her head"). In desisting, the teacher communicates that she is not "with it" by (*a*) desisting a minor deviancy while a major one is not acted upon (desisting two children for whispering while two others are throwing paper airplanes or chasing each other), or (*b*) desisting too late (after a deviancy has increased in seriousness or has begun to spread to other children).

Teachers were scored for Overlappingness on the basis of the percentage of their desists in which the teacher went all out to handle a deviancy and completely dropped the ongoing activity; they were scored for With-it-ness on the basis of the percentage of their desists that were aimed at the wrong target or came too late.

All correlations between teacher style and child behavior to be reported are between hypothesized good aspects of teaching style and desirable child behavior. For example, in setting up ranks for Spearman r's the classroom with the lowest deviancy rate would be ranked 1; the teacher with the highest With-it-ness would be ranked 1. A positive correlation between a teacher dimension and pupil deviancy or work involvement means that the teachers with more of the dimension had less deviancy and more work involvement.

Overlappingness, contrary to our expectations, did not correlate significantly with either deviancy or work involvement for either the E or Non E children. However, Overlappingness did correlate with With-it-ness ($r = .498$; $p < .01$, $N = 30$). With-it-ness, as can be seen in Tables 3 and 4, did correlate in many subsettings with

TABLE 3

SPEARMAN r'S BETWEEN THE WITH-IT-NESS OF TEACHERS AND THE DEVIANCY RATE OF CHILDREN

Subsetting	All grades				Grades 1 & 2			
	Nondisturbed children	N	Emotionally disturbed children	N	Nondisturbed children	N	Emotionally disturbed children	N
Seatwork in teacher's sphere	.439	14	.547*	14	—	3	—	3
Seatwork not in teacher's sphere	.386*	20	.517*	19	.559*	11	.709*	10
All seatwork	.480**	27	.524**	25	.568*	11	.685*	10
Subgroup recitation	.361	21	.432*	18	.321	11	.191	10
Class recitation	−.108	20	.327	18	−.136	6	.179	7
All recitation	.198	27	.413*	26	−.021	12	−.066	12

* $p < .05$; one-tailed test.
** $p < .01$; one-tailed test.

both deviancy and work involvement—the higher a teacher's degree of With-it-ness, the higher the work involvement of the children and the lower their deviancy rate. It should also be noted that correlations for E children were always in the same direction as for Non E children. In fact, the significant correlations obtained were either significant for the E children and not for the Non E children, or tended to be higher for the E than Non E children.

Since With-it-ness and Overlappingness correlate with each other, why does only With-it-ness correlate with children's behavior? Our current interpretation is that Overlappingness *enables* a teacher to receive information about what is going on, but that the children do not perceive or categorize this teacher dimension, or even notice it. What they do see and react to, however, is what the teacher *does* and communicates—she stops deviancy in time or she makes mistakes in picking deviancy targets when she is perceived by the children as knowing what is going on regarding their behavior, she has more surface managerial success,

TABLE 4

SPEARMAN r'S BETWEEN THE WITH-IT-NESS OF TEACHERS AND THE WORK-INVOLVEMENT RATE OF CHILDREN

Subsetting	All grades				Grades 1 & 2			
	Nondisturbed children	N	Emotionally disturbed children	N	Nondisturbed children	N	Emotionally disturbed children	N
Seatwork in teacher's sphere	.230	14	.766**	14	—	3	—	3
Seatwork not in teacher's sphere	.199	20	.293	19	.396	11	.697*	10
All seatwork	.227	27	.591**	25	.505	11	.727*	10
Subgroup recitation	.020	21	.534*	18	.182	11	.233	10
Class recitation	−.205	18	.335	18	−.657	6	.607	7
All recitation	−.042	27	.398*	25	.073	12	.448	12

* $p < .05$; one-tailed test.
** $p < .01$; one-tailed test.

that is, induces worklike behavior and restrains deviancy. Receiving or having knowledge is not sufficient unless the fact that this occurs is communicated to children.

One function of a teacher is to initiate activities and to maintain movement within them. Transition incidents should provide points where one could sample the manner in which a teacher manages the activity flow of a classroom. Transition points are those times when the children are instructed to stop an old activity ("Put spelling papers away.") and start a new one ("Get your arithmetic books out."). Old and new activities were distinct periods that teachers and pupils would presumably recognize and label in the same manner. They generally correspond to common content categories: arithmetic, spelling, reading group, etc.

All transition points were coded for the teacher's treatment of the activity to be stopped and of the activity to be started. The code included such items as the behavior mode of the children at the time of the transition; the presence of simple delineation or delineation with Clarity, Group Alerting, Challenge-Arousal, Zest; the attention property of the signal; the presence of antiresolution properties; the appropriateness of timing; and the children's immediate reaction to the induction.

Of particular relevance at this time are the following teacher behaviors, designated as antiresolution properties, that oppose her overt intent to have the children stop the old activity and begin the new one:

1. Providing contrary props, or passing out props that were judged to lure the children away from the teacher-designated official new activity—for example, passing out spelling tests the children had previously taken while telling the children to start their arithmetic.

2. Returning to old activity, or designating a new activity and then initiating an issue pertaining to the old—for example, asking how many children had all their spelling words right after having told them to stop spelling and start arithmetic.

3. Dangling, or signaling the pupils to start a new activity, and then dropping it in midair, so to speak, leaving the children with nothing to do.

4. Thrusting, or injecting a transition order at a time or in a manner indicating the teacher's unawareness of the children's receptivity to this message. For example, a teacher had been working with a reading group while the Robins were doing arithmetic seatwork. Suddenly, and in a manner indicative that her own intent was the only determinant of the timing—she had not looked up to see what the Robins were doing nor given verbal warning—the teacher thrust the order: "Robins, take out your reading books." A thrust is similar to a person butting in on the conversation of two other persons without either waiting for them to notice his presence or easing in by listening to what they are talking about.

5. Giving conflicting, confusing directions, or giving too many orders or contradictory things for the children to get with.

Teachers were scored and ranked on the basis of the percentage of their transitions that contained at least one of the above antiresolution properties.

The correlations between teachers' antiresolution scores during transition periods and children's behavior are presented in Tables 5 and 6. A positive correlation means that teachers with the lowest percentage of antiresolution techniques have the least deviancy or most work involvement in their class-

TABLE 5

SPEARMAN *r*'S BETWEEN TEACHERS' TRANSITION-RESOLUTION SCORES AND THE DEVIANCY RATE OF CHILDREN

Subsetting	All grades				Grades 1 & 2			
	Nondisturbed children	*N*	Emotionally disturbed children	*N*	Nondisturbed children	*N*	Emotionally disturbed children	*N*
Seatwork in teacher's sphere	.236	14	.227	14	—	3	—	3
Seatwork not in teacher's sphere	.347	20	.529*	19	.425	11	.627*	10
All seatwork	.308	27	.310	25	.425	11	.670*	10
Subgroup recitation	.412*	21	.530*	19	.527*	11	.282	10
Class recitation	.155	20	.336	18	.071	6	.357	7
All recitation	.350*	27	.386*	26	.397	12	.362	12

* $p < .05$; one-tailed test.

rooms. There are a sufficient number of significant correlations to justify the conclusion that one meaningful parameter of classroom management pertains to techniques of managing transitions. This dimension of classroom management, moreover, applies to both E and non E children in the same direction.

Having established a relationship between the techniques a teacher utilizes in directing children's movements from one activity to another and the degree to which she succeeds in inducing school-appropriate behaviors in these activities, one may well ask about the nature of the activities in which movement is programed.

One question about the programed activities in the school day has to do with the length of activities. A concept in common usage among teachers is that of "attention span"—the length of time that a child can remain involved in an activity before becoming "bored," restless, and desirous of quitting.

Having delineated transition points between one activity and another in all teachers' programs, the authors could measure the average length of activities by dividing the length of total

TABLE 6

SPEARMAN *r*'S BETWEEN TEACHERS' TRANSITION-RESOLUTION SCORES AND THE WORK-INVOLVEMENT RATE OF CHILDREN

Subsetting	All grades				Grades 1 & 2			
	Nondisturbed children	*N*	Emotionally disturbed children	*N*	Nondisturbed children	*N*	Emotionally disturbed children	*N*
Seatwork in teacher's sphere	.486*	14	.222	14	—	3	—	3
Seatwork not in teacher's sphere	.372	20	.474*	19	.380	11	.567*	10
All seatwork	.426*	27	.642**	25	.498	11	.600*	10
Subgroup recitation	.495*	21	.592**	19	.536*	11	.512	10
Class recitation	.255	18	.362	18	.057	6	.500	7
All recitation	.394*	27	.618**	25	.547*	12	.621*	12

* $p < .05$; one-tailed test.

** $p < .01$; one-tailed test.

time observed by the number of activities. Teachers were then ranked on the basis of the average length of time that they required the children to remain in an academic activity.

In no case did the authors find a significant correlation between average duration of academic activities and deviancy rate or work-involvement rate of children. Correlations were computed for activity length for both E and Non E children separately in the upper and lower grades and for all grades combined.

It must be concluded that attention span, when this concept is made equivalent to time as such, is not a significant aspect of activity programing in classrooms for either emotionally disturbed or nonemotionally disturbed children.

The authors then attempted to look at programing from the viewpoint of satiation theory. Satiation is defined as the process of decrease in liking for an activity due to repetition. In terms of the child-behavior variables measured in this study, it would be expected that increasing satiation would result in decreased work involvement and increased deviancy.

Experimental studies of satiation have demonstrated that felt progress and variety decrease the rate of satiation. The video tapes, being restricted to data about overt behavior, could not provide measures of how much progress a child felt. Conceivably, however, it could be possible to obtain measures of the amount of change and variety programed into the classroom day (or, in this case, half day).

The problem was to determine how different one activity is from another. Doing arithmetic is some change from doing spelling words. However, there are also similarities: children are sitting, writing, at the same seats, etc. It is also obvious that more change is involved in a transition from "seatwork spelling" to "recess" than in a transition from "seatwork spelling" to "seatwork arithmetic."

Accordingly, a nine-category coding system was designed to measure the amount of change between one activity and another: Content—19 categories such as reading, arithmetic, drawing, geography; Content Category—academic, semiacademic, and nonacademic; Teacher Presentation Pattern—eight categories such as lectures, questions; Props—from routine to unique; Group Configuration; Child Responsibility—such as listens passively, initiates own pace; Leadership—such as teacher, alone, another child; Overt Behavior Mode—four degrees from sedentary to physically vigorous activity; and Location—five subcategories, such as own seat, outside of room.

Each activity was coded for each of the nine major categories; change was scored at transition points. A change in any category was scored 1, and the number of categories scored as changed were summed to obtain a transition variety-change score. At any one transition a classroom could obtain a score of from 1–9. (A content change or teacher-sphere change was minimally required to designate a transition.) The sum of these scores was determined for each teacher. This score represents a combination of variety and contrast: The score is increased by increasing the variety of activities entered and by increasing the amount of difference between temporally adjacent activities.

No significant positive correlations were found between the amount of programed variety change in a teacher's half day and the degree of school-appropriate behavior of children. Of 64 correlations—no correlations were computed for "cells" containing fewer than 8 cases after being broken down by grade, setting, and E or Non E—

26 were negative, and 3 of these were significant at the .05 level—a finding that could be accounted for by chance.

One may conclude that the sheer amount of variety and change that a teacher programs in a classroom is not a dimension significant to managerial success.

This absence of a positive correlation was puzzling. Perhaps the "coffee-break" theory of handling problems of work satiation was pertinent. (Coffee breaks make the total work situation more tolerable, but do not actually change the attraction of the work as such; thus, the attractiveness of "tightening the bolt over and over" is not changed by the nonrelated coffee.)

Accordingly, a variety-change score was developed, based only on school-unique activities, namely, those activities that most children would perceive as concerned with learning or skill development, excluding the variety-change scores of entertainment, rest, physical exercise, recess, and nonacademic game activities.

The correlations between overall variety change for learning-related activities and children's behavior were higher than those obtained between children's behavior and variety change for all activities. (Of 64 pairs of correlations, 59 were higher.) And, except for a few negative correlations for the upper grades only, all were positive. However, significant positive correlations were obtained only in recitation subsettings, and only for all grades combined or for Grades 1 and 2 only. Taken separately, none of the correlations for the upper grades were statistically significant. For subgroup recitation settings, the following correlations were obtained in Grades 1 and 2 with deviancy rates: .72 for all children ($p < .01, N = 11$); .65 for Non E children ($p < .05, N = 11$); .59 for E children ($p < .05$, $N = 10$). With work-involvement rate, the correlations were .75 for all children ($p < .01, N = 11$); .75 for Non E children ($p < .01$, $N = 11$); and .67 for E children ($p < .05$, $N = 9$).

The above correlations support the theory that programing for variety change is a significant dimension of classroom management. It should be noted, however, that variety change *within* an activity was not scored. Would differentiating within-activity variety change provide a more refined analysis of teaching style?

To obtain refined variety-change scores for seatwork nonsphere settings, three refinements were added to the previous coding and scoring. One deals with the type of intellectual challenge required in the seatwork assignment; four categories were designated: (*a*) requires attention and perseveration only, such as copying words; (*b*) requires perseveration plus skill usage, such as doing addition of arithmetic numbers; (*c*) requires problem solution, such as solving word problems in arithmetic; and (*d*) requires creative effort, such as making up a story. The second refinement in the previous scoring deals with the designation of post-termination activities—"fillers." Does the teacher provide activities for the children when their assigned seatwork is finished, and, if so, how different are these activities from the rest of the seatwork? The authors reasoned that making available permissible school-type activities for children to engage in after required work is completed constitutes an antisatiation technique. A change score was thus computed between the last official seatwork task and the official filler, and added to the previous change score for seatwork activities.

Third, the variety-change score was refined by dividing the time spent in seatwork by the total variety-change score. This provided an index of seatwork variety change. The lower the index, the greater the experienced variety change in a period of time.

The results for this analysis are presented in Table 7. It can be seen that programed variety change within a subsetting is significantly related to the school-appropriate behavior of children in that subsetting, and that the direction of the relationship is the same for both E and Non E children. (SW Sp subsettings were not analyzed, as only three classrooms in Grades 1 and 2 had these subsettings. The development of a code for variety change in recitation subsettings has not yet been completed.)

Unlike the other teacher-style dimensions where the direction of the relationship with pupil behavior was the same for the lower and upper grades, the correlations of variety change are in opposite directions for the lower and upper grades.

Programing to reduce satiation is a significant dimension of teacher style, though this must be planned differently for younger than for older groups. It should be noted that, in the sample studied, the average variety-change index for *both* grade levels was 15 minutes. Since the duration of activities as such has not been found to be a significant variable, this difference between the lower and upper grades cannot be explained by the single statement that younger children have "shorter attention spans." One possibility is that "felt progress in learning" is a more important anti-satiation factor for the older children than for the younger children. Experienced variety may be relatively more salient for the younger children. The older children, then, might require activities of sufficient length to enable them to experience progress in learning and competence.

Programed variety change is also relevant to the problem of the disruptive effects of the deviant behavior of disturbed children. The correlations between variety change and contagion rates are in the same direction as those for deviancy and work involvement. For Grades 1–2 these are .658 ($p < .05$, $N = 10$) and for Grades 3–5, −.357 ($N = 7$). (There were only seven classes among the upper grades in which E children initiated a deviancy during SW.)

This study does indicate that dimensions of concrete teacher techniques can be delineated that make a difference in how children behave in classrooms. What is more, these are not unique for the emotionally disturbed

TABLE 7

SPEARMAN *r*'S BETWEEN SEATWORK VARIETY CHANGE AND CHILDREN'S BEHAVIOR IN SEATWORK

Grades 3-5								Grades 1-2							
Deviancy				Work involvement				Deviancy				Work involvement			
Emotionally disturbed children		Nondisturbed children		Emotionally disturbed children		Nondisturbed children		Emotionally disturbed children		Nondisturbed children		Emotionally disturbed children		Nondisturbed children	
Rate	*N*	Rate	*N*	Ratio	*N*	Ratio	*N*	Rate	*N*	Rate	*N*	Ratio	*N*	Ratio	*N*
−.413	9	−.667*	9	−.250	9	−.667*	9	.600*	10	.873**	11	.539*	10	.827**	11

* $p < .05$; one-tailed test.

** $p < .01$; one-tailed test.

children. In all cases, the direction of effect of a particular parameter of teaching style was the same for disturbed as for nondisturbed children.

One might also consider the implications of the findings of this study in relation to the training of teachers. For one thing, these findings point to the necessity of discovering dimensions of teaching style that are relevant to the ecology of the classroom and to a teacher's position in this setting. They justify a degree of skepticism about extrapolating dimensions of adult-child relations from other settings (homes, psychotherapy clinics) and applying these directly to teacher-child relations. They also raise questions about the fruitfulness of analyzing teachers on the basis of "personality" characteristics as compared to concrete techniques of programing activities and initiating and maintaining movement in the program. And, without the intent of minimizing the importance of studying individual children, the findings do suggest placing a higher priority on training for group management than is currently emphasized in educational psychology curricula.

REFERENCE

KOUNIN, J. S., & GUMP, P. V. The ripple effect in discipline. *Elementary School Journal,* 1958, 158–162.

(Received May 3, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 1, 14-17

RELATION OF ACHIEVEMENT MOTIVATION AND TEST ANXIETY TO PERFORMANCE IN PROGRAMED INSTRUCTION

HOWARD R. KIGHT AND JULIUS M. SASSENRATH
State University of New York, Buffalo *University of California, Davis*

The purpose of this study was to investigate the influence of achievement motivation and test anxiety on performance in programed instruction. A quasi-projective measure of achievement imagery, a test-anxiety questionnaire, and an achievement pretest were administered to 139 undergraduate pupils and related to their performance on programed instruction. Performance was analyzed using 3 criteria: (a) time needed to complete the material, (b) number of incorrect responses, and (c) a short-term retention test. The high-achievement-motivated students performed better on all 3 criteria and hence learned more efficiently with programed instruction than low-achievement-motivated students. High-test-anxiety students also worked faster and made fewer errors than low-anxiety students, but failed to exhibit higher retention scores.

Considerable attention has been given in recent years to the study of achievement motivation and its concomitant effect upon classroom learning. McClelland (1951) and his followers have spurred much of this interest by providing ways to objectively measure or identify achievement motivation. Yet the research up to this point has not produced any clear-cut agreement with regard to the influence of achievement motivation within the classroom.

When grade-point averages (GPAs) are employed, as they often are, to represent scholastic achievement, the disparity becomes even clearer. Studies, for example, by McClelland, Atkinson, Clark, and Lowell (1953) and by Weiss, Wertheimer, and Groesbeck (1959) have reported moderate correlations of .51 and .34, respectively, between achievement motivation and GPAs of subjects (*Ss*), using college students as *Ss*. Other studies by Lowell (1952) and Parrish and Rethlingshafer (1954), for example, have been unable to show any significant relationship between grades and achievement motivation. Admitting that GPAs represent a rather general estimate of academic achievement, it could easily be argued that some other means of assessing achievement is needed.

Frequently some type of paper-and-pencil task has been employed as a measure of achievement—ranging from simple code substitution to highly complex arithmetic problems. The results generally have shown that *Ss* with high achievement motivation tend to exhibit higher performance scores than do low-achievement-motivated *Ss*, especially on the simpler tasks (French, 1955). And although paper-and-pencil tasks may afford a more reliable index of achievement, they do not usually represent actual class material. If the influence of achievement motivation is to be fully understood, particularly at the college level, it would seem highly desirable to have the tasks represent actual course material.

One technique which seems valuable in this regard is programed instruction. Not only is a student's rate of performance easily recorded, but also the difficulty of all items (tasks) is kept intentionally low. High-achievement-motivated students thus have an opportunity to work faster and to show higher overall proficiency than do *S*s with low achievement motivation. One of the two major purposes of this study, then, is to verify the hypothesis that high-achievement-motivated *S*s will exhibit higher performance on programed instruction than will *S*s with low achievement motivation.

A second major purpose of this study is to investigate the relationship of a second, equally important motivation variable—test anxiety—to performance on programed instruction. Although the research interest in test anxiety seems to be increasing, its influence upon performance within the classroom, particularly on programed instruction, has not been sufficiently studied.

The results have generally shown that low-anxious *S*s perform better than high-anxious *S*s when the task involves either complex tasks or stress-motivating instructions (Lucas, 1952; Nicholson, 1958; Sarason, Mandler, & Craighill, 1952). Considerably less attention has been devoted to ascertain situations whereby high-anxiety *S*s would actually excell in performance. One of the few studies to approach this problem was by Grimes and Allinsmith (1961) in which it was found that highly anxious children worked best in highly structured situations. Since programed instruction does represent a fairly well-structured, nonstressful situation as a rule, it would seem plausible to expect the high-test-anxious *S*s to work faster and show greater proficiency than low-test-anxious *S*s on this type of task. This, then, represents the second major hypothesis to be verified by this investigation.

METHOD

The *S*s for this study included 139 undergraduate students in educational psychology at Indiana University. Before beginning the programed-instruction phase of the course, all *S*s were administered the following three instruments:

1. *Iowa Picture Interpretation Test* (Form RK). This test, developed by Hedlund (1953), consists of four alternative responses to each of 24 Thematic Apperception Test (TAT) pictures. *S* is instructed to rank order each set of four alternatives on the basis of how well he describes what is happening in the picture. *S*'s achievement motivation, labeled his "achievement imagery" score, is derived by summing the ranks assigned by him to each alternative previously designated as an A_I choice.
2. *Test Anxiety Questionnaire* (TAQ). Mandler and Sarason (1952) developed this instrument to assess a person's feelings and attitudes about a variety of testing situations, for example, achievement, aptitude, etc. The strength of a person's subjective response is marked along a 0–15 point continuum and a total score obtained by summing for all 35 items.
3. *Achievement pretest.* This test, consisting of objective-type questions, was designed to serve as a preliminary measure of *S*'s skill and knowledge of basic principles related to test construction and evaluation. The questions closely paralleled but did not duplicate statements appearing on the programed-instruction material.

During the 3 weeks which followed, approximately 2 hours each week, *S*s were given a series of three linear-type programed booklets, dealing with the construction and analysis of classroom achievement tests.[1] Special care was taken at the outset to explain the main purposes and advantages underlying this type of instruction and to stress the relevance of the material being presented. Each booklet contained about 50 supply-type frames to which *S* responded on a specially prepared answer sheet. A record was kept of the number of correct responses for each *S* as well as the number of minutes spent. Mimeographed reading material, unrelated to the programed material, was handed to each *S* as he completed each unit so that all *S*s began a unit at the same time and spent the same number of class periods. The experimenter supervised all aspects of the programed instruction.

DESIGN

On both the achievement-motivation and test-anxiety scales, *S*s were categorized as being either High (above the median) or Low. *S*s were also classified as High (above the median) or Low on the basis of their pretests so that within each level of achievement motivation and test anxiety *S*s could be matched according to their pretest ability. This produced a 2 × 2 × 2 factorial design and permitted the application of a three-way analysis of variance on each of several performance measures.

---

[1] The authors are indebted to John Krumboltz of Stanford University for providing most of the programed instruction material.

To adequately represent *S*s' performances on the programed instruction, three separate measures were obtained. These included: (*a*) length of time each *S* took to complete the material, (*b*) number of frames answered incorrectly, and (*c*) *S*'s proficiency on an achievement test[2], administered at the conclusion of the programed instruction.

RESULTS

Table 1 presents a comparison of means on the three performance measures for *S*s classified as High or Low in achievement motivation and High or Low in test anxiety.

TABLE 1
MEAN SCORES ON THREE MEASURES OF PERFORMANCE IN PROGRAMED INSTRUCTION

Groups	Time	Errors	Retention
Low AM–Low TA	125.2	26.5	20.1
Low AM–High TA	124.6	19.9	20.6
High AM–Low TA	107.6	20.2	24.1
High AM–High TA	100.0	13.8	27.0

Note.—AM: Achievement Motivation. TA: Test Anxiety.

In order to statistically test the two hypotheses—namely, that a higher degree of motivation, either achievement motivation or test anxiety, would produce better performance on the programed instruction—an analysis of variance was performed on each of the three performance measures.[3]

On the first performance measure—minutes taken to complete the programed material—all three main effects were significant and in the predicted direction. *S*s who were high in achievement motivation, test anxiety, or who scored high on the pretest tended to work through the program at a faster rate (Fs = 11.20, 5.08, 7.10, respectively; $df = 1/131$). It should be noted, however, that a significant interaction occurred between achievement motivation and test anxiety ($F = 4.16$; $df = 1/131$).

On the second performance measure—number of incorrect responses on the programed material—two main effects were significant. *S*s high in achievement motivation or test anxiety tended to commit fewer errors (Fs = 9.25, 5.18, respectively; $df = 1/131$). There was also a significant interaction effect between test anxiety and pretest ability ($F = 4.08$; $df = 1/131$).

As a third measure of *S*s' performance on the programed instruction, scores on the short-term retention test were analyzed. Two main effects of achievement motivation and pretest ability were significant. Thus, *S*s high on either of these two variables scored reliably higher on the retention test than did *S*s low on the same variable (Fs = 8.00, 37.50, respectively; $df = 1/130$).

CONCLUSIONS

The results of the present investigation substantially confirm both hypotheses. Specifically, it has been demonstrated that groups with high achievement motivation or high test anxiety (*a*) required less time to complete the programed learning material, (*b*) made fewer overt errors on the material, and (*c*) received higher scores on a short-term retention test than did low-achievement-motivated or low-test-anxiety groups. These findings are in accord with some of the reported research, though it is hazardous at this point to make any direct comparisons. The occurrence of several significant interactions adds further scepticism and doubt regarding the overall effects of any one motive.

---

[2] The achievement test was constructed as a parallel form of the pretest. Both tests were found by a previous analysis to have internal consistency reliabilities above .75 and to intercorrelate .81 with each other.

[3] These and succeeding calculations were performed on the IBM 709 data processing system at the Indiana University Computing Center. Grateful acknowledgement is made to Dale Hall who made computer time available for this study.

One important conclusion which has stemmed from the analysis of both Time and Errors scores is that on well-structured but relatively easy material such as programed instruction, the highly motivated *S*s work much more efficiently than the less-motivated *S*s. In summary, the high-achievement-motivated and high-test-anxiety groups work faster and make fewer errors than those who are low on either variable. One possible explanation for this is that the highly motivated *S*s are able to derive greater benefit from the immediate knowledge of results, characteristic of programed instruction.

REFERENCES

FRENCH, E. G. Some characteristics of achievement motivation. *Journal of Experimental Psychology,* 1955, **50,** 232–236.

GRIMES, J. W., & ALLINSMITH, W. Compulsivity, anxiety, and school achievement. *Merrill-Palmer Quarterly,* 1961, **7,** 247–269.

HEDLUND, J. L. Construction and evaluation of an objective test of achievement imagery. Unpublished doctoral dissertation, State University of Iowa, 1953.

LOWELL, E. L. The effect of need for achievement on learning and speed of performance. *Journal of Psychology,* 1952, **33,** 31–40.

LUCAS, J. D. The interactive effects of anxiety, failure, and intra-serial duplication. *American Journal of Psychology,* 1952, **65,** 59–66.

MANDLER, G., & Sarason, S. R. A study of anxiety and learning. *Journal of Abnormal and Social Psychology,* 1952, **47,** 166–173.

McCLELLAND, D. C. *Personality.* New York: Sloane, 1951.

McCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.

NICHOLSON, W. M. The influence of anxiety upon learning. *Journal of Personality,* 1958, **26,** 303–319.

PARRISH, J., & RETHLINGSHAFER, DOROTHY. A study of the need to achieve in college achievers and non-achievers. *Journal of Genetic Psychology,* 1954, **50,** 209–226.

SARASON, S. D., MANDLER, G., & CRAIGHILL, P. G. The effect of differential instructions on anxiety and learning. *Journal of Abnormal and Social Psychology,* 1952, **47,** 561–565.

WEISS, P., WERTHEIMER, M., & GROESBECK, B. Achievement motivation, academic aptitude and college grades. *Educational and Psychological Measurement,* 1959, **19,** 663–665.

(Received January 21, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 1, 18-22

CULTURAL INFLUENCES ON RETENTION OF LOGICAL AND SYMBOLIC MATERIAL[1]

HERBERT HARARI AND JOHN W. McDAVID
University of Miami

American educational and training practices are characterized predominantly by objective testing based on recognition whereas those of Israel are characterized by subjective testing based on reconstruction. To explore effects of these cultural differences upon retention, a logical content task and a content-free task comprised of symbols were administered to 320 12-year-old Ss, equally divided according to culture (American, Israeli), sex, recall interval (immediate, 48-hour delayed), and retention measure (recognition, reconstruction). Significantly higher reconstruction scores for the Israeli group partially support the hypothesis that cultural factors may give rise to differences in performance on retention tasks. The results suggest the relevance of experiential factors to individual differences in performance on various types of learning measures.

Early anthropological studies, introspective and anecdotal rather than experimental in nature, suggested that memory is primarily a function of the intrinsic interests of the individual (e. g., Bartlett, 1932; Nadel, 1937). These interests were presumed to be related to the individual's past meaningful experiences within his social group. Only one really experimental cross-cultural study dealing with memory has been reported: a comparison of immediate-recall ability of Caucasian, Japanese, and Chinese children in Hawaii (Louttit, 1931). However, this study suffers some inadequacy since all subjects (*S*s) belonged to the same educational system as well as to fairly homogenous subcultures.

In America, experimental studies of memory dealing with recall and recognition have rather consistently shown scores on recognition measures to be superior to scores on recall measures (e. g., Anastasi, 1932; Burt & Dobell, 1955; Luh, 1922; Phillip & Peixotto, 1949; Stalnaker, 1935). Experimental cross-cultural studies have focused their attention on various aspects of individual and group performance on tests of intelligence rather than of memory (Buros, 1959). Thus, there is support for Klineberg's (1954) contention that "in absence of research on memory in a variety of cultures, it is not possible to state whether the 'laws' of learning and forgetting would be equally valid under all social conditions [p. 226]."

The study reported here compared the performance of American and Israeli children on various memory tasks using both recognition and reconstruction as measures of retention. The experience of the senior author in dealing with high school students in both American and Israeli educational institutions led to the assumption that the academic training of American children relies heavily on recognition methods in assessing learning while that of Israeli children relies heavily on reconstruction methods. Thus it was expected that these cultural variables which differentiate the educational background of the two groups should facilitate performance on recognition measures for American students, and conversely, should facilitate performance on reconstruction measures for Israeli students.

---

[1] The data on which this report is based were collected in connection with the first-named author's master's thesis in psychology at the University of Miami (Harari, 1962). The constructive suggestions of Robert B. Tallarico as well as the assistance of Pinhas Sommer in collecting data in Tel Aviv are gratefully acknowledged.

METHOD

Subjects

*S*s were 320 12-year-old children, 160 boys and 160 girls. Of these, half were enrolled in middle-class suburban public schools in southern Florida (the American sample), and half were enrolled in a comparable school in Tel Aviv (the Israeli sample). All *S*s were Jewish. *S*s were drawn at random from the total populations of each school and were assigned randomly to each of the experimental conditions described below.

Learning Tasks

A meaningful content task, the Marble Statue test, was adopted from Whipple's (1915) manual after being judged as culture fair by three American and three Israeli educators. A nonmeaningful symbol task was specially constructed after considerable pretesting. It consisted of five sets of two-digit numbers with a simple geometrical design opposite each set. Black digits and symbols approximately 3″ × 3″ were presented on a white 22″ × 24″ card. Both the symbol task and the content task were used to secure immediate as well as delayed measures of retention.

Retention Measures

Although the same learning tasks were presented to all *S*s, two methods of assessing retention were utilized. Half the *S*s in each cultural sample (40 *S*s of each sex) were told to reproduce the Marble Statue story in writing and to reproduce the symbols with their matching digits in writing. The remaining half were administered a one-out-of-three multiple-choice recognition test on each task. The recognition test on the Marble Statue task contained items dealing with 67 distinct ideas in the story as identified by Whipple (1915); in the five-item recognition test on the symbol task, the digits served as stimuli, and the geometric designs as multiple-choice alternatives. Within each of the subgroups thus generated, half the *S*s underwent immediate assessment of retention, and the other half were tested after a 48-hour delay. *S*s were not told that they would later be tested on what they had learned and were thus left to assume that the experiment was terminated after the first session. No special attempt was made to prevent communication among *S*s during the 48-hour delay interval, but there was no evidence in either group that the experimental situation excited any appreciable amount of discussion among the students during this time. Although the testing was conducted by different experimenters in Tel Aviv and in Florida, explicit and formalized procedures were outlined, and each experimenter adhered literally to these standardized instructions. In order to allow use of analysis of variance procedures, a fixed number of *S*s were scheduled for inclusion in each experimental condition. Thus, four experimental subgroups, each containing 20 males and 20 females, were erected within each cultural group: immediate recognition, immediate reconstruction, delayed recognition, and delayed reconstruction.

Scoring

In scoring reproduction of the content task, one point was allowed for each correct idea included as prescribed by Whipple's (1915, p. 336) manual. The recognition test was scored by allowing one point for each correct answer. In the symbol task, the reconstruction measure was scored by allowing two points for each correct matching of properly reproduced symbols and digits and one point for any other correct reproduction of either a symbol or a set of digits. Since the measures of recognition and reconstruction were not inherently comparable, no direct comparisons of the two scores within each cultural subgroup were possible.

RESULTS AND DISCUSSION

Mean scores for each of the experimental subgroups on each of the measures of retention are reported in Tables 1 and 2.

Reconstruction scores for the content task were submitted to a three-dimensional (Sex of *S* × Cultural Group × Delay Interval) factorial analysis of variance. Scores for the Israeli students were significantly superior to those for the American students ($F = 30.16, p < .01$). Scores for both cultural groups on the immedi-

TABLE 1

SUMMARY OF SCORES FOR AMERICAN AND ISRAELI STUDENTS ON IMMEDIATE AND DELAYED RECONSTRUCTION AND RECOGNITION MEASURES OF RECALL WITH A MEANINGFUL CONTENT LEARNING TASK

Group	Males		Females	
	M	*SD*	*M*	*SD*
American				
Reconstruction measures				
Immediate recall	26.85	6.35	24.85	8.70
Delayed recall	20.20	5.70	22.05	4.85
Recognition measures				
Immediate recall	6.55	.80	6.50	.66
Delayed recall	5.75	.94	6.50	.66
Israeli				
Reconstruction measures				
Immediate recall	34.85	6.71	31.30	6.70
Delayed recall	24.90	6.50	26.10	6.45
Recognition measures				
Immediate recall	6.65	.57	6.90	.30
Delayed recall	5.70	1.00	6.80	.46

ate-recall measure were superior to scores on the delayed-recall measure ($F = 33.92$, $p < .01$). Although the overall simple sex factor was not significant, the interaction between sex of *S* and recall interval was significant ($F = 4.12$, $p < .05$). While this interaction has no bearing on the purpose of the investigation, it is attributable to sex differences in opposite directions within each of the recall-interval groups: Boys were superior to girls on the immediate-reconstruction measure but were inferior to girls on the delayed-reconstruction measure. This pattern occurred within both cultural groups.

Scores on the recognition measure of recall for the content task, as well as those on both the recognition and

TABLE 2

SUMMARY OF SCORES FOR AMERICAN AND ISRAELI STUDENTS ON IMMEDIATE AND DELAYED RECONSTRUCTION AND RECOGNITION MEASURES OF RECALL WITH A MEANINGLESS SYMBOL LEARNING TASK

Group	Males		Females	
	M	*SD*	*M*	*SD*
American				
Reconstruction measures				
Immediate recall	12.20	3.20	12.30	1.25
Delayed recall	4.70	1.85	7.65	5.40
Recognition measures				
Immediate recall	4.60	.58	4.70	.74
Delayed recall	3.30	1.65	4.30	.57
Israeli				
Reconstruction measures				
Immediate recall	13.05	1.85	12.45	2.05
Delayed recall	10.20	4.55	11.00	3.60
Recognition measures				
Immediate recall	4.50	.57	4.80	.40
Delayed recall	4.30	.78	4.80	.31

reconstruction measures for the symbol tasks, were distributed inappropriately and required nonparametric analysis. Using the Mann-Whitney *U* statistic (Siegel, 1956), no significant differences between the two cultural groups were revealed. Thus, the hypothesis that American students should perform better on recognition measures found no direct support.

However, a comparison of delayed-recall performance measures for each of the cultural groups on the symbol-association task showed that Israeli students were superior to their American counterparts on both reconstruction ($p < .01$) and recognition ($p < .01$) scores. While it is *possible* that the Israeli students were simply intellectually superior to the American students, there is no basis either from the overall data in this study or from characterization of the two *S* populations to support such an assumption. Rather, it appears that students who do well in reconstruction tasks may do well also in tasks involving recognition but that the reverse does not necessarily hold true. Assuming no real differences in actual potential or ability between the two cultural groups, these observations are interpreted as likely consequences of differences in training and experience with particular educational and testing practices.

If it were possible to compare scores on the reconstruction and recognition measures directly, a more satisfactory direct conclusion with reference to the initial hypotheses of this study could be drawn. Unfortunately, however, there was no means of equating scores on the two measures. Furthermore, it was not possible to utilize standard score conversions for such a comparison since the score distributions on the recognition measures were highly skewed away from the ceiling of the measure. These score distributions suggest that the recognition measure (which had been constructed on the basis of pretesting with American students) was perhaps too easy, resulting in a stacking of scores at the ceiling of its score range.

Further exploration of cultural differences with respect to various methods of assessing retention certainly seems warranted. The data presented here provide only tentative support for the initial hypotheses advanced, and considerable refinement of the assessment methods utilized in further studies will be necessary in order to provide definitive evidence concerning those hypotheses. Nevertheless, these data strongly suggest an effect of experiential and training factors upon individual differences in performance on various measures utilized in assessing the retention of learned material. The role of such "test-taking skills" is not confined merely to achievement testing but may also have serious implications with respect to cross-cultural comparisons of intelligence and aptitude measures.

REFERENCES

ANASTASI, A. Further studies on the memory factor. *Archives of Psychology,* 1932, **142,** 60.

BARTLETT, F. C. *Remembering.* London: Cambridge Univer. Press, 1932.

BUROS, O. K. *The fifth mental measurements yearbook.* Highland Park, N. J.: Gryphon Press, 1959.

BURT, H. E., & DOBELL, E. M. The curve of forgetting for advertised material. *Journal of Applied Psychology,* 1955, **9,** 5–21.

HARARI, H. Cultural influences on retention of logical and symbolic material. Unpublished master's thesis, University of Miami, 1962.

KLINEBERG, O. *Social psychology.* New York: Holt, 1954.

LOUTTIT, C. M. Racial comparison of ability in immediate recall of logical and nonsense material. *Journal of Social Psychology,* 1931, **2,** 205–215.

LUH, C. W., The conditions of retention.

6·3·81

Psychological Monographs, 1922, **31**(3, Whole No. 142).

NADEL, S. F. A field experiment in racial psychology. *British Journal of Psychology*, 1937, **28**, 195–211.

PHILLIP, B. R., & PEIXOTTO, H. R. Recall and recognition of nonsense syllables. *American Journal of Psychology*, 1949, **62**, 228–237.

SIEGEL, S. *Nonparametric statistics*. New York: McGraw-Hill, 1956.

STALNAKER, J. M. Recognition and recall in vocabulary test. *Journal of Genetic Psychology*, 1935, **46**, 463–464.

WHIPPLE, G. M. *Manual of mental and physical tests: II*. Baltimore: Warwick & York, 1915.

(Received February 10, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 1, 23–26

EFFICIENCY IN MEASURING THE EFFECT OF ANXIETY UPON ACADEMIC PERFORMANCE[1]

NEIL A. CARRIER AND DONALD O. JEWELL[2]
Southern Illinois University *Valparaiso University*

The extent to which academic examination performance could be predicted from several measures of academic achievement anxiety was tested by 3 studies. To attain the greatest (most efficient) predictive utility from the shortest amount of testing time, interaction effects between anxiety measures were determined, expressed as cross products, and included as independent variables in each multiple-regression equation. The dependent variable was score on a final examination in introductory psychology. This procedure increased predictive efficiency using selected anxiety measures. Prediction of examination performance was better for female than male Ss.

Recent research on the effect of anxiety upon performance has seen the development of a number of specific instruments for the measurement of anxiety related to academic achievement performance (Alexander & Husek, 1962; Alpert & Haber, 1960; Mandler & Sarason, 1952; Sarason, 1958). Alpert and Haber (1960) have concluded that such scales are better predictors of academic performance than are general anxiety scales and seem to account for variance in such performance beyond that accounted for by aptitude measures. Preliminary unpublished work by the authors, primarily performed within analysis of variance designs, not only supported the first of these conclusions, but also suggested interactions between specific anxiety measures in accounting for variance in course examination performance.

Since one of our objectives was to be able to test the accuracy of predictions of academic performance made from anxiety scores, it was decided to employ multiple-correlation procedures in subsequent work. In multiple-regression analyses possible interaction effects may be determined by including the cross products of variables as independent variables and testing the significance of their contribution to the multiple-correlation coefficient. If such a procedure could thus identify cross products which would increase a multiple correlation beyond that based upon anxiety scores alone, a greater "efficiency" of prediction would be possible in the sense of accounting for more variance with the same amount of testing time. In addition to this empirical rationale, the position was taken that since a number of different types of anxiety measures were to be employed, each presumably measuring a somewhat different aspect or kind of anxiety, there is some common-sense logic in assuming that certain combinations of scores may mean more than the independent scores alone. As a consequence, three consecutive studies were conducted to determine what minimum combination of several measures of test anxiety and their cross products would provide the best prediction of final-examination performance.

---

[1] The support of the Graduate School, Southern Illinois University, in the conduct of this research is gratefully acknowledged.

[2] Now at Southern Illinois University.

The authors shared equal responsibility for the planning and conduct of the research and the preparation of this report. The assistance of William Vicars in the preparation of data during the early phases of the project is acknowledged.

STUDY I

Procedure

The subjects were 137 students in an introductory psychology course at Southern Illinois University. Approximately at midquarter they took Alpert and Haber's (1960) 19-item Achievement Anxiety Test (AAT). From two subsets of items "debilitating anxiety" (A−) and "facilitating anxiety" (A+) scores were obtained. Immediately prior to a second hourly examination, a 33-item form of the Anxiety Differential (AD) was given (Alexander & Husek, 1962). A final examination of 125 multiple-choice questions provided the measure of academic performance. A multiple-regression analysis was carried out using the final-examination raw score as the criterion variable, and raw scores on the AD, both scales of the AAT, and the three possible cross products of these scores as independent variables. Each independent variable was tested to determine the significance of its contribution in accounting for variance in the criterion variable. The "best" mutiple-regression equation (largest multiple-correlation coefficient from the fewest independent variables) was sought for each sex and for the sexes combined.

Results

The Study I portion of Table 1 shows the first-order and multiple-correlation coefficients of the three anxiety measures (alone and in various combinations) with the criterion measure. The most notable feature of the data is the fact that adding their cross products to the three anxiety measures, as independent variables in the multiple correlation, increased the coefficient. This is more obvious in the case of the female than in the male subjects.

These results were regarded as encouraging enough to pursue a second study. The same measures and techniques employed in the first study were used to permit a check on the reliability of the results obtained in the first study. The unique aspect of the second study was the addition of more anxiety measures.

STUDY II

Procedure

The following quarter a second group of 151 subjects was tested the same way as in Study I. In addition, Sarason's (1958) Test Anxiety Scale (TAS) was given about midquarter and scored for the usual six scales. After determining the contribution of each of the new variables and the cross products in accounting for variance in the criterion (final-examination performance), two scales

TABLE 1

FIRST-ORDER AND MULTIPLE CORRELATIONS OF CERTAIN VARIABLES WITH FINAL-EXAMINATION RAW SCORE

Correlations	Study I			Study II			Study III		
	M $N = 98$	F $N = 39$	M & F $N = 137$	M $N = 94$	F $N = 57$	M & F $N = 151$	M $N = 33$	F $N = 33$	M & F $N = 66$
First-order									
A+	.23*	.22	.24**	.19	.42**	.26**	.13	.56**	.37**
A−	−.22*	−.19	−.23**	−.37**	−.41**	−.38**	−.32	−.57**	−.45**
AD	.07	.11	.06	−.18	−.13	−.16*	−.31	−.08	−.19
TA				−.32**	−.43**	−.35**			
NA				−.26*	−.03	−.20			
Multiple									
A+,A−	.26**	.24	.27**	.37**	.47**	.39**	.32	.64**	.50**
A+,A−,AD	.32**	.29	.31**	.37**	.49**	.39**	.38	.65**	.51**
A+,A−,TA				.38**	.49**	.40**			
A+,A−,NA				.40**	.51**	.39**			
A+,A−,AD,(A+)(A−), (A+)(AD),(A−)(AD)	.38**	.49*	.39**	.39*	.60**	.39**	.43	.68**	.56**
A+,A−,TA,(A+)(A−), (A+)(TA),(A−)(TA)				.43**	.58**	.43**			
A+,A−,NA,(A+)(A−), (A+)(NA),(A−)(NA)				.46**	.68**	.46**			

* $p \leq .05$.
** $p \leq .01$.

of the TAS were chosen as promising: Test Anxiety (TA) and Need for Achievement (NA).

Results

The Study II portion of Table 1 shows selected first-order and multiple-correlation coefficients. The results obtained in the first study using the three scales (A+, A−, AD) were fairly generally repeated in the second study. Notable increases in coefficients were apparent, however, in the case of the female subjects. Also, it was again seen that adding cross products as independent variables in the regression equation increased the multiple-correlation coefficient. The same sex difference was observed, with prediction likely to be better for females than for males.

Since the best combination of three anxiety measures (and their cross products) seemed to be the A+, A−, and NA scales, these were selected for the third study in which (*a*) confirmation of the results obtained here was to be sought; and (*b*) these three variables (and their cross products) were to be used to predict final-examination performance.

STUDY III

Procedure

A third group of 66 subjects was administered the AAT and the NA scale during the fourth week of a quarter. The latter was administered as a separate instrument rather than as a subset of items of the TAS. This was done in the interest of economy of time in testing. The AAT and NA scale were administered together, in that order, and took a maximum of 15 minutes for subjects to complete. The basic objective was to account for as much of the variance in final-examination performance as possible with as brief an assessment of anxiety as possible.

Using multiple-regression equations developed for each sex from the data of the second study, predicted final-examination scores were calculated using the six scores obtained from the A+, A−, NA, and their three cross products. At the end of the quarter the actual score on the 125-item multiple-choice final examination was obtained. The first-order correlations of predicted with actual score were .32 for males and .52 for females. The size of these coefficients is encouraging when it is remembered that the data upon which they were based were obtained from only 15 minutes of nonintellective testing, and that the predictive equations were based upon data obtained from a relatively small sample of the student population.

The third portion of Table 1 shows the first-order and multiple correlations on the data collected in the third study. The results from the second study were generally supported, but again the size of the female subjects' correlations showed an apparent increase and were substantially higher than those of the male subjects. The contribution to the multiple correlation by the cross products of the three anxiety scales is less for females in the third study than in the second. This is probably due to the unexpectedly high multiple correlation obtained using only the raw scores as independent variables. Since there are probably a number of other variables not included here that are related to academic performance (aptitude, etc.), very high multiple correlations that omit them probably cannot be expected. It appears that as the multiple correlation based upon these anxiety raw scores approaches .70, the contributions of cross products necessarily diminish.

Discussion

The data of these studies support the contention that academic examination performance can be predicted from scores on self-report measures of anxiety. This prediction is increased when such scores are selectively combined in the form of a multiple-regression equation. Even further efficiency may be obtained from the measures by the selective inclusion in the regression equation of the cross products of the measures.

The present data are in disagreement with the findings of Dember, Nairne, and Miller (1962) which led them to conclude that the AAT "is not as useful for female as for male subjects [p. 428]." The present data support an opposite interpretation. While it is not known what differences in sample characteristics or other variables are responsible for this discrepancy, an unpublished study by the first-named author is suggestive. That study found a sex discrepancy in first-order correlations betweeen the AAT scales and hourly examination scores consistent with that reported by Dember et al. (1962). Contrary to the present study, however, subjects were drawn from two "small" sections of a course in child psychology, rather than from the larger introductory course taught (and examined) in the large lecture format. These groups apparently resemble those tested by Dember et al. more than either resembles those of the present study.[3]

In addition to the AAT and NA scale, the AD and TA scales appear to have some merit in accounting for examination performance. A methodological procedure which may have prevented the AD scale from making its maximum contribution in the first two studies was the fact that this measure of "situational anxiety" was not given immediately before the final examination from which the criterion scores were obtained. However, it was given in each study in conjunction with a previous hourly examination, and earlier data gathered by the authors showed AD scores correlated about the same with examinations taken immediately following its administration and those taken several weeks later.

The findings also suggest a number of interesting questions. If anxiety factors do play a considerable role in performance on academic tests, how much of the variance in the scores on college admission tests is a result of these factors? What nonanxiety personality variables interact with anxiety to influence examination performance? Do the effects of anxiety upon performance vary with the type of examination (recall, recognition, etc.)? What situational conditions affect examination performance, and can the detrimental effects be reduced? The authors are now pursuing studies addressed to these questions.

REFERENCES

ALEXANDER, S., & HUSEK, T. R. The Anxiety Differential: Initial steps in the development of a measure of situational anxiety. *Educational and Psychological Measurement*, 1962, **22**, 235–348.

ALPERT, R., & HABER, R. N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, **61**, 207–215.

DEMBER, W. N., NAIRNE, F., & MILLER, F. J. Further validation of the Alpert-Haber Achievement Anxiety Test. *Journal of Abnormal and Social Psychology*, 1962, **65**, 427–428.

MANDLER, G., & SARASON, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, **47**, 166–173.

SARASON, I. G. Interrelationships among individual difference variables, behavior in psychotherapy, and verbal conditioning. *Journal of Abnormal and Social Psychology*, 1958, **56**, 339–344.

(Received February 17, 1965)

---

[3] The unpublished study also found the AAT scales (especially the A−) correlated higher with performance on the essay than with the multiple-choice questions in the course.

Journal of Educational Psychology
1966, Vol. 57, No. 1, 27-34

RELATIONS OF SOCIAL DATING ATTRACTIONS TO ACADEMIC ORIENTATION AND PEER RELATIONS

ROBERT E. GRINDER

Department of Educational Psychology, University of Wisconsin

The present study investigates the hypothesis that relatively strong interest in the rewards of social dating and involvement in high school responsibilities clash detrimentally. Social dating questionnaires consisting of 4 Likert scales (Sexual Gratification, Independence Assertion, Status Seeking, and Participative Eagerness) and requesting a variety of personal data were administered to 393 boys and 346 girls in Grades 10, 11, and 12. Academic performance, as measured by grade-point average, was negatively associated with interest in all 4 aspects of dating for boys and sexual and independence-assertion attractions for girls. Number of friends was significantly associated with boys' interest in all aspects of dating investigated. Clique membership was reliably associated for both sexes with the status-seeking, and for girls with the sexual, aspect of dating.

The youth-culture literature suggests that adolescents who lean heavily upon peer functions for support are likely to be diverted from attaining adult-approved objectives of socialization (Allport, 1961; Parsons, 1962; Wilfert, 1963). One might expect, therefore, that adjustment to high school responsibilities would be affected adversely by the extent to which adolescents strive for the attractions of the youth culture. The assumption is highly general, and the present study attempts to evaluate empirically the degree to which it holds for adolescents in the 10th, 11th, and 12th grades by comparing interest in certain specific attractions of social dating with various aspects of orientation toward high school and toward peer relations.

The belief that youth-culture and adult goals may be incompatible stems traditionally from psychoanalytic theory, which has provided a basis for the widely adopted belief that participation in the youth culture serves primarily as an outlet for adolescent rebelliousness (Hurlock, 1955; Keniston, 1962; Pearson, 1958). Analyses of youth-culture activities also have emerged recently through the influence of social learning theory, which stresses the formative influence of reward-and-punishment features of familial and cultural functions (Bandura, 1964; Elkin & Westley, 1955; Parsons, 1962). This viewpoint asserts that adolescents may be more interested in participating in activities of the youth culture than of families or schools to the extent that its attractions and privileges are more appealing. The theory implies that adolescents who seem defiantly independent (rebellious) of adult-sponsored goals may be attempting to preserve peer status and to maintain a flow of rewards from youth-culture cohorts. Hence, both psychoanalytic and social learning theory lead to the belief that extensive interest and involvement in youth-culture attractions may be associated with unsatisfactory adjustment to high school responsibilities. Moreover, the social learning view stresses that youths' diversion from their obligations may be, in part, a result of the powerful appeal of youth-culture rewards, satisfactions, and comforts.

The present investigation of youth-culture functions is restricted, in order that it might be reduced to manageable proportions, to the appeal of social dating. For present purposes, *social dating is defined as encompassing all situations in which the principal actors constitute a heterosexual pair, are unmarried, and are unchaperoned, free agents*. Dating patterns, ranging from casual dating to engagement, encompassed by the definition offered here appear to share several common attractions (cf. Method, below). The rewards most easily obtained through social dating are available in various contexts and in countless other configurations; as defined here, social dating merely is one of many systems through which they are obtained. However, analysis of the dating social system provides an excellent realm for investigating the youth culture because maturation of sexual propensities and cultural emphases upon heterosexual associations are likely to make the rewards of social dating especially compelling for most youth.

The interpretations drawn from both psychoanalytic and social learning theory and the analysis of social dating presented above suggest that adolescents strongly involved in the attractions of dating would be least dedicated to high school responsibilities. The rewards for social dating are promptly and tangibly acquired from peers, and their consistent attainment requires investment of considerable time; by and large, the permanent benefits of discharging high school obligations, that is, preparing oneself for additional education and citizenship, entail postponement of gratification, intangibleness, and consumption of vast amounts of time. The two systems apparently compete for youths' attention, and many youths seemingly forsake one for the other. Specifically, then, one might expect that orientation toward the incentives of social dating would be inversely related to grade-point average, attention to homework, and vocational aspirations. Since school extracurricular activities are adult sponsored, controlled, and reflect essentially adult-oriented rewards, one would also expect an inverse relationship to prevail between the social dating interests and participation in extracurricular activities such as school clubs and councils. Further, inasmuch as peers are the primary administrators of youth-culture rewards, adolescents who seek the attractions of dating should be those who have a coterie of friends.

METHOD

Subjects

Subjects were 393 boys and 346 girls from the 10th, 11th, and 12th grades, drawn in fairly equal numbers from "college," "general," and "basic" curricular groupings of an urban public high school, which enrolled some 2,000 pupils who were widely diversified in terms of social class, academic skills, and vocational aspirations.

Questionnaire[1]

A comprehensive list of the attractions offered by social dating was compiled primarily from the descriptive interpretations of sociologists, the replies of several adolescents when queried as to why they dated, and from various articles in newspapers and teenage magazines. Analysis of the data from these sources led to the postulation of the following four incentive categories:

1. *Sexual Gratification* (SG). Dating offers sanctioned opportunities to make physical contact with members of the opposite sex, for example, to dance cheek to cheek, to hold hands at the theater, to stroll with an arm around the consort's waist, to caress, to neck

---

[1] Inasmuch as space limitations prohibited extensive discussion of the questionnaire, a 20-page document containing complete copies of the revised forms for both sexes, means, standard deviations, and biserial *r*'s showing the item-criterion correlations of each item may be obtained by writing the author.

in a parked car or in some secluded place, etc.

2. *Independence Assertion* (IA). Dating provides a context in which to achieve autonomy of adult authorities and of generally accepted standards of society. Drag racing on lonely streets, fudging one's age in order to gain admittance to an "adults-only" movie, and dating members of the opposite sex whose reputations are tarnished are among the possibilities. Dating may also be a vehicle for deviating from family religious practices, disdaining family political beliefs, ignoring parental strictures on curfew hours and sex-conduct rules, and being heedless of family propriety in crossing of social-class lines, in smoking, and in drinking.

3. *Status Seeking* (SS). Dating offers occasions for associating with prestigious members of the opposite sex. One may prefer consorts who are champions or leaders in debating, athletics, or dramatics. Dating may be a vehicle for entering an "in group." Dating partners may be sought who are excellent models for developing social etiquette, grace, and sophistication.

4. *Participative Eagerness* (PE). Rewards notwithstanding, dating sometimes is appealing for the sake of the activity it provides, for example, obtaining an identifiable reward may be of less concern than avoiding loneliness, boredom, anxiety, work-role responsibilities, and functions with parents or same-sex peers.

Each of the above categories provided the basis for an unidimensional Likert-type scale. Separate scales, employing similar, sex-appropriate items, were prepared for each sex. Each scale is comprised of items or statements that purport to describe a choice that a same-sex protagonist has made between two discernible alternatives: one, the attraction of social dating, and the other, either a desirable nondate goal or adherence to cultural norms, mores, parental strictures, etc. Each item re-creates the situational dynamics of a double approach-avoidance conflict (Miller, 1944). The subjects (*S*s) are asked to evaluate the protagonist's behavior in each item—whether or not they have ever been in the situation—on the basis of five degrees of agreement and disagreement. The summation of each *S*'s responses to the items of the separate scales yields for each category a total score, which is interpreted as representing an indication of *S*'s motivation for that particular category of social dating incentive (Selltiz, Jahoda, Deutsch, & Cook, 1962).

In constructing the items, an attempt was made to provide simple and realistic conflicts, to avoid response set by reversing the direction of the items, and to minimize arousal of self-consciousness and defensiveness. Also, efforts were aimed at writing median items in order to maximize variance and, very importantly, to provide a safeguard against including items that adolescents—and school officials who must sanction them before they can be administered—might regard as bizarre and/or irrelevant.

Initially, 15 items for each scale were randomized for presentation in a 60-item booklet and were pilot tested upon two high school populations of 220 and 406 *S*s, respectively. After the second revision, each scale was reduced to 10 items. The final form of the questionnaire, therefore, comprised 40 items. An example of boys' *Sexual Gratification, Independence Assertion, Status Seeking,* and *Participative Eagerness,* respectively, is given below:

1. A boy and girl are parked in his car beside the lake. She is shy but they have necked before. He would like to neck again, but she hints about leaving. He tries to change her mind.

2. A couple have stayed out once or twice beyond the girl's curfew hour, as set by her parents. They are at an interesting party and they decide to stay longer since her parents don't enforce the curfew.

3. A boy is eager to make the "honor" service society, but he must know more of the top members. Being a member will make him a school leader, too. He dates an interesting and attractive girl who is a member, *just* so he will be invited to the society's parties—thus he will get to know the members and be asked to join the society.

4. A famous sports star is coming to town to make a public appearance for one night only. A boy is eager to see him. He knows his buddies are going and he plans to go with them. But then he finds out that a girl he has wanted to date would like him to ask her out to another show. He gets the date.

Table 1 shows the intercorrelations among the four scales for each sex. The dimensions under study were not expected to be entirely unrelated, and forcing the independence of the scales further seemed unwarranted; moreover, the existence and nature of the relationships are themselves worthy of scrutiny. The correlation between the Status Seeking and Participative Eagerness scales for girls is .40 whereas the corresponding comparison for boys is .12, and it may be that participation itself represents acquisition of status for girls. The differences among

TABLE 1

INTERCORRELATIONS AMONG SEXUAL-GRATIFICATION (SG), INDEPENDENCE-ASSERTION (IA), STATUS-SEEKING (SS), AND PARTICIPATIVE-EAGERNESS (PE) SCALES FOR BOYS AND GIRLS

Scale	Boys ($N = 393$)				Girls ($N = 346$)			
	SG	IA	SS	PE	SG	IA	SS	PE
SG	.87[a]	.31	.16	.22	.81[a]	.32	.21	.31
IA		.76[a]	.16	.22		.78[a]	.24	.35
SS			.58[a]	.12			.64[a]	.40
PE				.70[a]				.66[a]

[a] Diagonal entries are Hoyt reliability values.

the reliability measures reported in the diagonals of Table 1 reflect in part differences among the items for each of the scales. The SS items for example, are relatively more dissimilar than the SG items, and the phenomenon probably mirrors the real-life scene; for example, *S*s are less likely to diffuse the focus of sexual symbols and inclinations than of status symbols.

Personal Information

The personal information items were grouped into three categories: *validity,* including such items as frequency of dating, age beginning dating, curfew hour, nights out per week, and access to car; *academic,* including grades received in school, hours spent studying, and academic degree aspirations; *peer relations,* including number of close friends, participation in high school activities, and membership in cliques. Each of the personal information items was followed with a list of appropriate responses, one of which was to be checked.

Procedure

*S*s were administered the questionnaire and personal information items in their classrooms by one of four experimenters, none of whom was a teacher. *S*s were asked to raise their hands upon completing the questionnaire; then, as each *S* finished, he was handed the personal information items, and his questionnaire was collected. The two parts together consumed approximately 30 minutes. Only the most general instructions were given to *S*s. They were told that all the data would be the confidential property of the university researchers; moreover, *S*s were explicitly told not to put their names on the papers. However, in order that the two sets might be matched, they were asked to use their birth dates. Serious cooperation on the part of *S*s was stressed. All *S*s were cooperative, and most appeared to be highly interested in participating in the project.

Statistical Analysis

Major statistical evaluation in this study was accomplished through use of a two-way analysis of variance design. Unequal frequencies in the subclasses necessitated the use of a special procedure described by Scheffé (1959). Each of the four scales of the questionnaire was treated as a dependent variable, and the personal information items were employed as classificatory factors in a univariate analysis of variance. Three sets of analyses were computed for each sex. Frequency of dating was cross-partitioned with each of several validity measures; additionally, frequency of dating and grade in school, respectively, were cross-partitioned with each of the academic and peer-relations variables.

RESULTS AND DISCUSSION

The results presented in Tables 2 and 3 include only main effects since none of the interactions with the frequency-of-dating and grade-in-school classifications were significant. Further, a main effect for the grade-in-school classification, by 10th, 11th, and 12th grades, did not appear, and these data also are not reported.

The data offered in Table 2 pertain mainly to construct validation. One would expect, for example, that adolescents expressing relatively greater interest in the attractions represented by the scales would participate more frequently in the dating system and would be more likely to possess opportunities with which to pursue the rewards. The findings of Table 2 confirm the expectation in respect to frequency of dating. Mean scores on the scales increase as dating becomes more active. The main effect is significant for boys on the four scales, that is, SG ($p < .001$), IA ($p < .001$), SS ($p < .05$), and PE ($p < .001$); for girls on the

TABLE 2

MEAN SCORES AND RESULTS OF ANALYSES OF VARIANCE FOR SEXUAL GRATIFICATION (SG), INDEPENDENCE ASSERTION (IA), STATUS SEEKING (SS), AND PARTICIPATIVE EAGERNESS (PE) IN TERMS OF FREQUENCY OF DATING AND OTHER ASPECTS OF SOCIAL DATING

Antecedent	Boys					Girls				
	N	SG	IA	SS	PE	*N*	SG	IA	SS	PE
Frequency of dating										
No dates yet	81	27.4	24.7	24.9	27.9	48	25.4	25.5	23.7	25.3
1 per wk.	123	32.0	29.0	26.1	29.1	96	28.0	25.8	23.9	27.3
1 per mo.	106	34.2	31.0	27.2	31.0	97	29.8	27.0	23.6	26.4
3 + per wk.	83	36.7	32.5	27.0	32.4	105	32.8	29.7	24.1	27.3
F ratio		27.46***	24.30***	3.50*	10.68***		14.11***	5.86***	*ns*	*ns*
Age dating began										
13 or younger	85	35.9	32.5	28.1	32.4	45	31.6	28.4	23.1	26.3
14	99	34.2	31.3	26.4	30.4	106	30.2	28.2	24.1	27.9
15	74	34.0	29.8	26.6	30.2	101	30.2	27.0	23.3	26.5
16 or older	54	32.4	28.1	25.9	30.0	46	29.8	26.6	23.6	26.7
F ratio		*ns*	4.95**	*ns*	*ns*		*ns*	*ns*	*ns*	*ns*
Curfew										
12 P.M.	119	33.8	29.5	27.1	30.5	190	30.0	27.1	23.8	27.0
2 A.M.	72	35.3	31.3	25.6	30.6	50	32.3	29.7	23.0	27.3
Anytime	121	34.4	32.8	27.1	31.4	58	29.4	27.1	23.3	26.7
F ratio		*ns*	5.20**	*ns*	*ns*		*ns*	3.50*	*ns*	*ns*
Nights allowed out per wk.										
1	60	31.7	27.7	26.3	28.9	69	30.9	27.7	25.4	28.3
2–3	177	34.6	31.2	27.1	30.8	206	30.9	28.0	23.6	27.1
4 or more	75	36.2	33.2	27.1	32.5	23	32.4	27.2	22.7	30.2
F ratio		4.92**	9.58***	*ns*	4.46*		*ns*	*ns*	*ns*	*ns*
Car availability										
Always	130	35.1	32.4	26.7	30.5	162	30.4	27.7	23.1	27.2
Sometimes	119	34.0	30.4	26.7	31.3	92	30.5	27.7	24.4	26.6
Seldom/never	63	32.9	28.0	26.9	30.8	44	29.2	26.3	23.7	26.4
F ratio		*ns*	11.04***	*ns*	*ns*		*ns*	*ns*	*ns*	*ns*

* $p < .05$.
** $p < .01$.
*** $p < .001$.

SG ($p < .001$) and IA ($p < .001$) scales. It is noteworthy that the Status-Seeking and Participative-Eagerness scales yielded nonsignificant main effects for adolescent girls on the frequency-of-dating classification. Girls who actively date and girls who seldom date appear either equally eager or equally disinclined to pursue the rewards of the dating system. The image of the present population of adolescent girls as being reluctant to engage in dating is palpably incongruent with the facts of contemporary life, and it would seem that the appropriate interpretation is that all girls, whether they date frequently, are eager to participate in the system and to seek the peer status it proffers.

The findings presented in Table 2 offer scant support for the notion that opportunity to acquire the attractions of dating is associated with enhanced interest in them. Possession of freedom to participate in the dating system, however, appears to exert a differential effect upon the two sexes. In the instance of every classificatory variable shown in Table 2, the main effect on the Independence-Assertion scale is significant. Inspection of the means shows that given more or less unrestrained liberty, boys rather than girls seem more likely to assert their individuality in the context of the dating system. It appears also that boys' zeal for participating in dating and for seeking physical gratifications are associated with the number of evenings they are allowed out per week, but not by whether they enjoy an early or late curfew or by whether they have access to a car. Of what avail, after all, are a late curfew and a car if opportunities to capitalize upon these advantages are highly limited? On the whole, girls who have relatively greater autonomy and opportunity would also be expected to be especially interested in the attractions of dating since, by and large, the inducements are there for the taking. As in the instance of frequency of dating, however, impositions and restrictions notwithstanding, girls appear equally zealous to acquire the rewards of the dating system.

The belief that youths' enthusiasm for the incentives of social dating would be associated inversely with academic performance is supported by the data presented in Table 3, Part I. The means based upon grade-point average show that high scores on the dating scales tend to be associated with low grades. The main effect is significant for boys on each of the four scales (SG, $p < .05$; IA, $p < .05$; SS, $p < .01$; PE, $p < .001$), and for girls on the SG ($p < .05$) and IA ($p < .05$) scales. Two measures associated with academic performance corroborate the interpretation. The data of Table 3, Part I, also show that the means of the dating scales that showed significant main effects decreased for both sexes as study time and aspirations increased.

The data presented in Table 3, Part II, suggest, as expected, that an inverse relationship exists between participating in adult-sponsored extracurricular activities and certain of the social dating attractions. For the extracurricular-participation classification, involving school clubs and councils, significant main effects appeared for boys on the SS ($p < .05$) and PE ($p < .001$) scales, and for girls, on the IA ($p < .05$) scale. Also, the expectation that adolescents' interests in dating may be associated with peer interaction, on the assumption that peers are the primary administrators of youth-culture rewards, was largely confirmed. On the number-of-friends classification, the analyses of variance yielded a significant main effect for boys on each of the four scales (SG, $p < .001$; IA, $p < .001$; SS, $p < .001$; PE, $p < .05$). The mean scores for each scale increased with the possession of a greater number of friends. Tests for the number-of-friends measure were nonsignificant for girls. However, when the girls who reported having three or more friends are separated in terms of whether or not they share secrets with one another, following Hollingshead's (1949) analysis of clique behavior, the expectation receives some support. The main effects are significant for the SG ($p < .05$) and SS ($p < .001$) scales, and the discrepancies between the means indicate that the girls with close-knit peer relations scored higher on the scales. This comparison holds only for boys on the SS scale ($p < .05$).

To the extent that the variables employed in the present study discrimi-

TABLE 3

MEAN SCORES AND RESULTS OF ANALYSES OF VARIANCE FOR SEXUAL GRATIFICATION (SG), INDEPENDENCE ASSERTION (IA), STATUS SEEKING (SS), AND PARTICIPATIVE EAGERNESS (PE) IN TERMS OF ACADEMIC AND PEER-RELATIONS VARIABLES

Antecedent	Boys					Girls				
	N	SG	IA	SS	PE	*N*	SG	IA	SS	PE
I. Academic variables										
Grade-point average										
A	77	31.4	27.2	25.4	29.1	116	27.3	25.3	23.5	26.1
B	99	31.8	28.7	25.9	28.1	91	30.9	27.1	23.2	27.0
C	147	32.9	29.7	26.5	30.9	102	31.3	29.1	24.1	27.6
D	70	34.6	32.1	28.2	32.2	37	29.9	28.1	24.8	26.2
F ratio		2.62*	6.82*	4.16**	5.54***		3.61*	3.46*	*ns*	*ns*
Hrs. of study at home per wk.										
3–4	154	34.5	31.4	26.8	32.6	74	32.5	30.4	24.1	28.2
4–6	141	31.7	28.9	25.9	28.9	132	29.3	27.3	23.9	26.8
7+	98	30.3	26.5	25.3	28.5	140	28.5	25.6	23.4	26.1
F ratio		10.84***	16.88***	*ns*	18.54***		8.46***	13.38***	*ns*	3.38***
Aspirations										
High school degree	92	33.5	30.7	27.2	31.4	76	32.8	29.1	22.8	28.3
Vocational school	89	32.8	30.1	26.7	31.3	95	30.2	28.3	24.0	27.5
College	212	32.0	28.9	25.9	29.2	175	28.1	25.8	24.0	25.8
F ratio		*ns*	3.09*	*ns*	2.99*		10.50***	6.64**	*ns*	5.27**
II. Peer-relations variables										
Extracurricular participation										
0	135	32.8	30.2	27.2	31.7	68	29.5	28.4	22.5	26.7
2–3 per mo.	138	32.5	29.2	26.3	30.1	173	30.3	27.8	23.9	27.6
3+ per wk.	120	32.4	28.6	25.3	28.6	95	29.2	25.8	24.2	26.2
F ratio		*ns*	*ns*	4.10*	8.38**		*ns*	3.23*	*ns*	*ns*
No. friends										
0	51	29.6	26.1	25.2	28.9	64	29.1	27.0	23.6	26.2
1–2	136	32.2	29.1	25.6	29.7	91	30.4	28.7	24.7	27.4
3–4	108	32.0	28.9	26.2	29.8	97	29.5	26.7	23.8	26.4
5+	98	35.0	31.6	28.3	31.7	94	29.2	26.9	23.1	26.4
F ratio		6.91***	9.35***	5.88***	3.41*		*ns*	*ns*	*ns*	*ns*
Clique membership										
yes	151	34.0	30.7	27.8	31.1	159	29.9	27.1	23.9	27.0
no	55	32.1	28.9	25.7	30.6	32	27.6	26.0	21.2	26.0
F ratio		*ns*	*ns*	4.97*	*ns*		4.32*	*ns*	10.83***	*ns*

*$p < .05$.
**$p < .01$.
***$p < .001$.

nate among adolescents in terms of their interest in obtaining the rewards of the social dating system, the data of the present investigation suggest that, in addition to the negative, punitive, frustrating aspects of school life that may propel adolescents away from school, there may exist inducements that entice them away. Moreover, it appears that the various attractions of social dating are equally influential across the 10th, 11th, and 12th grades. On the whole, poor academic performance is consistently related to interest in sexual gratification and in asserting independence. Nearly every adolescent is imbued with the cultural imperative to find a mate, and it seems that adolescents who have been sidetracked from fulfilling school responsibilities find solace in indulging, to a degree, in both sexual gratifications and refractoriness, the rewards for which are peer-culture administered. Yet, for boys at least, peer relations and the status-seeking and the participative features of dating appear to be the more closely associated. Boys who are nonparticipants in extracurricular activities seem especially interested in participating in dating and in using the dating system for seeking status. Sex differences also emerge in respect to peer relations. Having friends per se is an important correlate of boys' motivation for all the resources of dating, whereas having close, intimate friends appears to be the more critical variable for girls.

REFERENCES

ALLPORT, G. W. Values and our youth. *Teachers College Record,* 1961, **63,** 211–219.

BANDURA, A. The stormy decade: Fact or fiction? *Psychology in the Schools,* 1964, **1,** 224–231.

ELKIN, F., & WESTLEY, W. A. The myth of adolescent culture. *American Sociological Review,* 1955, **20,** 680–684.

HOLLINGSHEAD, A. B. *Elmtown's youth.* New York: Wiley, 1949.

HURLOCK, ELIZABETH B. *Adolescent development.* New York: McGraw-Hill, 1955.

KENISTON, K. Social change and youth in America. *Daedalus,* 1962, **91,** 145–171.

MILLER, N. E. Experimental studies of conflict. In J. McV. Hunt (Ed.), *Personality and the behavioral disorders.* Vol. 1. New York: Ronald, 1944. Pp. 431–465.

PARSONS, T. Youth in the context of American society. *Daedalus,* 1962, **91,** 97–123.

PEARSON, G. H. J. *Adolescence and the conflict of generations.* New York: Norton, 1958.

SCHEFFÉ, H. *The analysis of variance.* New York: Wiley, 1959.

SELLTIZ, CLAIRE, JAHODA, MARIE, DEUTSCH, M., & COOK, S. W. *Research methods in social relations.* New York: Holt, 1962.

WILFERT, O. Youth-adult relationships: The social milieu. *International Journal of Adult and Youth Education,* 1963, **15,** 43–49.

(Received May 10, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 1, 35–47

EFFECTS OF TEACHER AND PEER SUBCULTURES UPON STUDENT ASPIRATIONS[1]

DONALD L. THISTLETHWAITE AND NORMAN WHEELER
Vanderbilt University

A panel of 1,772 students enrolled at 140 colleges was surveyed at the beginning, middle, and end of college. Multiple-regression equations were developed for predicting terminal dispositions to seek advanced graduate or professional training from 8 precollege characteristics. Residual scores, used as indexes of changes in dispositions, were correlated with student reports of their college environments, and factor analyses of the latter were performed. 14 of 33 college press scales, and 3 of 9 factors were significantly correlated with both behavioral and aspirational residuals of men, while none of the scales or factors were significantly correlated with both kinds of residuals of women. The disposition to seek advanced training was manifested by an increasing number of panel members with increased exposure to college.

Previous studies of the effects of college upon aspirations to seek graduate or professional degrees have been limited to two modes of analysis. In the first, which may be called a "college productivity" analysis, the undergraduate institution is regarded as the unit of sampling, and an attempt is made to describe the characteristics of institutions which are unusually "high" or "low" with respect to their output of students who later obtain advanced degrees or other scholarly rewards (Astin, 1961, 1962b; Knapp & Goodrich, 1952; Thistlethwaite, 1959a, 1959b). The most frequently expressed methodological caution in recent productivity studies is the necessity of controlling differences in student quality among institutions so that comparisons on the basis of the attainments of graduates will not merely reflect diversities in entering classes. While there is suggestive evidence that both the structural and psychological characteristics of college environments influence motivation for advanced degrees, none of the positive studies have controlled more than a few college-input variables.

The second frequently used mode of analysis may be called an "analysis to find out who changes."[2] In this paradigm the student is regarded as the unit of sampling, and an attempt is made to correlate individual differences with respect to change in degree aspirations with actual or reported experiences during college (Astin, 1962a, 1963a, 1963b; Thistlethwaite, 1960, 1962a, 1962b, 1963). The present study, which follows the second paradigm, attempts to control as many as eight precollege characteristics, and thereby to expose hypotheses concerning the effects of teacher and peer subcultures upon student degree aspirations to more searching probes than have yet been reported.

---

[1] The research reported herein was supported through the Cooperative Research Program of the Office of Education, United States Department of Health, Education, and Welfare. In addition the early phases of research were partially supported by National Science Foundation, Old Dominion Foundation, and Ford Foundation grants to the National Merit Scholarship Corporation. We are grateful to Richard L. Gorsuch for consultation in the analysis of data.

[2] This term has been borrowed from Hovland, Lumsdaine, and Sheffield (1949, p. 316), who first used it to refer to a method of studying the relation of film effects to population variables.

METHOD

Sample

Members of the sample were first identified as juniors in secondary school when they participated in an aptitude test used to select scholarship recipients in a national scholarship competition. Examinees are known to differ from nonparticipating high school students in that they more frequently planned to enter college, more often had parents who wanted them to go to college, and more frequently took college preparatory courses (Dirks & Anderson, 1960). Only the 300,000 examinees having test scores equivalent to a national percentile rank of 64 or higher were designated for sampling. A 10% random sample, stratified by state, was drawn, and the first mail survey was sent to selected students in 1959, a few weeks before they would normally have entered college. Returns were received from 26,771 respondents, who comprised approximately 90% of the designated sample. There were proportionately fewer returns from students residing in the South and from students with lower aptitude-test scores. The bulk of the data from the first survey was analyzed for a purpose which does not concern us here.

The second mail survey was sent only to selected panel members who had enrolled in one of the 140 most popular colleges and universities attended by respondents to the initial survey. Names of respondents enrolling at these colleges were stratified by college of enrollment, and random samples of 60 names were selected from each stratum containing more than 60 names, while all names were selected from each stratum containing between 35 and 60 names. These procedures yielded a designated sample of 7,464 examinees to whom a second mail survey was sent in 1961, when these students would normally have completed the second year of college. Usable returns were received from 58%, or 4,330, of this group. Response rate varied by sex and aptitude level: Women responded more frequently than men, and students with high aptitude-test scores responded more frequently than those with low test scores.

The final survey was made in 1963, a few weeks before panel members would normally have entered graduate or professional school. It was limited to students who indicated in 1961 that they were proceeding through college in normal progression. Returns were received from 2,919, or 70%, of the designated sample. Three response biases were present in the returns: First, high-aptitude students tended to respond more frequently than low-aptitude students; second, there were proportionately more returns from students with higher grade-point averages as reported at the end of the second year of college; third, response rate was positively related to the student's degree aspirations as reported at the end of the second year of college.

The data to be reported are based only on panel members who participated in each of these surveys, and the cumulative response biases clearly resulted in the selection of a panel more disposed to seek advanced training than a typical sample of college graduates. These biases precluded any attempt to estimate from the panel the proportion of all college students who continue their training in graduate or professional schools, but they were not a source of invalidity in our attempt to test hypotheses concerning factors influencing student motivation for such training.

Criterion Questions

Included in the final survey were two items designed to provide indicants of motivation to seek advanced training: The first, hereafter referred to as a measure of *aspirations,* asked the respondent to indicate the highest level of education he expected to complete (e.g., BA, MA, first professional, or PhD degree); the second, hereafter called the *entry* criterion, asked the respondent to indicate whether or not he was enrolling immediately in a graduate or professional school. Responses to the aspiration question were scored with weights of 0, 1, and 2 for BA, MA, and professional or PhD degrees, respectively; responses to the entry question were scored with weights of 0 and 1 for nonentry and entry, respectively. Since it usually was not possible for nongraduates to enter graduate or professional school at this time, only data for the 1,970 panel members who had fulfilled all the requirements for graduation from college on or before September 1963 were included in the analyses to be reported.

Adjusted Criterion Measures

Precollege Characteristics. In an earlier analysis of data from the first two surveys sex and initial degree aspirations were the 2 predictors (among 11 on which reasonably complete information was available) most highly related to degree aspirations after 2 years of college. However, subgroups matched with respect to sex and initial degree aspirations but sorted according to changes in aspirations after 2 years of college differed with respect to several precollege characteristics. For example, subgroups showing increases in degree aspirations tended to have higher aptitude-test scores and more often had applied for freshman scholarships. Also the mothers and fathers of these students had higher levels of education and somewhat higher family incomes. Probable major field of study as reported at the beginning of college was also highly related to the student's intermediate degree aspirations. Accordingly the following precollege characteristics were controlled in this analysis:

1. Sex.
2. Degree aspirations at the beginning of college (1959).
3. National Merit Scholarship Qualifying Test (NMSQT) score in 1958.
4. Father's educational level.
5. Mother's educational level.
6. Number of freshman scholarship applications in 1959.
7. Family financial resources as reported in 1959.
8. Probable major field of study as reported in 1959.

Since sex and probable major field of study were nominal classifications, data for these attributes were recorded as "dummy variables" in the manner described by Tatsuoka and Tiedeman (1963, p. 148). Complete information on these precollege attributes was available for 1,772 of the panel members who had completed college by September 1963.

Total Regression Analysis. Multiple-regression equations were developed for predicting both criteria for the 1,772 panel members from the eight precollege predictors. Residual scores (discrepancies between actual and predicted criterion scores) were used as the principal dependent variables. These residuals may be interpreted as indexes of the extent to which each student *changed* his motivation for advanced training from the level predictable from his status on the selected precollege variables.

Standardization and Cross-Validation Analyses. Astin (1963a) used an actuarial method to achieve similar control over precollege characteristics, except that he derived his actuarial formula from a standardization sample and then applied it to a cross-validation sample. The following additional analyses were performed partly to investigate the equivalence of residuals based on "optimal" and "nonoptimal" regression equations and partly to analyze the effects of college upon men and women separately.

Random samples of male and female panel members (475 men, 412 women) were drawn for *standardization* samples. Data for these respondents were used to develop multiple-regression equations for predicting both the aspiration and entry criterion from the precollege characteristics. (Sex was omitted since it was controlled by performing separate analyses by sex.) Discrepancies between predicted and obtained criterion scores were calculated and are hereafter referred to as *standardization* residuals. The foregoing multiple-regression equations were then applied to the data for the remaining respondents (461 men, 424 women) who comprised the *cross-validation* sample. Errors of estimate, based upon applying these nonoptimal equations to subjects (*S*s) in the cross-validation samples are referred to as nonoptimal residuals. In addition, the data for cross-validation *S*s were used to develop new multiple-regression equations for predicting each criterion. Discrepancies between predictions based upon the latter equations and the corresponding obtained criterion scores may be termed optimal residuals. By correlating nonoptimal and optimal residuals for *S*s in the cross-validation sample, it was possible to study the extent to which the two kinds of residuals were equivalent.

Independent Variables

One class of measures of college environments consisted of *college press scales* containing questionnaire items on which the student was asked to describe the demands, expectations, or activities most characteristic of teachers and students at his college. These scales may be subdivided according to whether they called for description of the student's environment as a lowerclassman or as an upperclassman.

Lower-Class Press Scales. Each of these scales, administered to panel members at the end of their second year of college, contained 10 dichotomous items, 5 keyed so that a "True" response was weighted positively, and 5 keyed so that a "False" response was weighted positively. The "objects" of description were presumably collegewide, since students were *not* instructed to limit their reports to their own teachers or to their close friends. Unlike Stern's (1958) scales, each of these scales was devised so that it called for descriptions of faculty or students, but never both.

Upper-Class Press Scales. Each of these

TABLE 1

College Press Scales

	Scales and representative items
Faculty press (lower-class scales)	
1–1	Affiliation. (The faculty keeps pretty much to itself and rarely fraternizes with students.) F
1–2	Directiveness. (Assignments are usually clear and specific, making it easy for students to plan their studies effectively.) T
1–3	Enthusiasm. (The outstanding traits of the faculty are their dedication to and enthusiasm for their work.) T
1–4	Achievement. (The professors are really skillful at getting students to work to the limit of their capacities.) T
1–5	Compliance. (The faculty usually demands strict compliance with all course requirements.) T
1–6	Supportiveness. (Teachers are sensitive to student complaints and grievances and frequently try to remedy the situation.) T
1–7	Humanism. (All students here are expected to take numerous elective courses in English, literature, history, or philosophy.) T
1–8	Independence. (The college rarely encourages students to undertake independent projects or theses.) F
1–9	Vocationalism. (The college offers many really practical courses designed to prepare the student for his occupation.) T
Student press (lower-class scales)	
1–10	Competition. (The competition for high achievement is intense.) T
1–11	Estheticism. (There is a lot of interest here in poetry, music, painting, sculpture, architecture, etc.) T
1–12	Reflectiveness. (Students here have very little interest in the analysis of value systems and the relativity of societies and ethics.) F
1–13	Social Conformity. (Ignorance of social customs is no handicap to the student here.) F
1–14	Intellectualism. (Students here rarely get excited about a campus speaker.) F
Faculty press (upper-class scales)	
2–1	Positive Evaluations of Ability. (Their evaluations of my academic performance convinced me that I had a flair for course work in this area.) T
2–2	High Academic Standards. (Typically their courses demanded excellent academic performance by all students.) T
2–3	Independent Thinking. (They did little to help the student develop his imaginative and creative capacities.) F
2–4	Enthusiasm for Intellectual Values. (They clearly conveyed to me a sense of what it means to become deeply involved in a discipline or subject.) T
2–5	Advanced Training. (They encouraged students to do graduate work.) T
2–6	Convergent Thinking. (They typically demanded that answers to test questions be strictly to the point and answer the specific question asked.) T
2–7	Adequacy as Positive Role Models. (On the whole I am grateful to them for showing me a way of life worthy of imitation.) T
2–8	Excellence of Faculty Teaching. (Their lectures clearly revealed they had put a great deal of energy and thought into doing a good job of teaching.) T
2–9	Availability of Faculty to Students. (There was no opportunity to interact with them outside of our respective roles as "students" and "faculty.") F
2–10	Affiliation. (Their counseling and guidance of me was really personal, patient, and extensive.) T
2–11	Compliance. (They typically demanded strict compliance with all course requirements.) T
2–12	Structuring of Major Field of Study. (Their courses suggested to me that there was more disorder than order in their art or area of knowledge.) F

Table 1—Continued

Scales and representative items

Student press (upper-class scales)

2–13 Intellectualism. (We shared with each other the excitement of intellectual discoveries.) T
2–14 Unfavorable Self-evaluations. (My associates were more talented and better equipped to pursue their studies than I.) T
2–15 Cohesiveness. (It was easy to get a group of them together for card games, talk sessions, going to the movies, etc.) T
2–16 Rebelliousness. (They occasionally plotted some sort of escapade or rebellion.) T
2–17 Playfulness. (We rarely had much time for play or recreation.) F
2–18 Status. (Among my associates it was important socially to be in the right club or group.) T
2–19 Opposition to Faculty Influence. (Most of them felt their teachers had helped them to achieve greater direction, force, and clarity.) F
2–20 Openness to Experience. (They often went to concerts and art exhibits when they were available.) T

Note.—The item correlating highest with each scale is shown in parentheses. The T or F indicates whether agree or disagree responses, respectively, received greater weight in scoring the given item.

20 scales, which were included in the final survey sent to students a few weeks before they would normally have entered graduate or professional school, consisted of 8 items, 4 keyed so that "Agree" responses were weighted positively, and 4 keyed so that "Disagree" responses were weighted positively. Each item called for Likert-type ratings of degree of agreement on a 5-point rating scale. In completing these items respondents were asked to describe undergraduate teachers in their *major field of study* with whom they had taken most of their courses, or to describe student colleagues with whom they had associated most commonly. While some of these scales overlapped substantially in item content with corresponding lower-class press scales, most of them were designed to represent new dimensions of college press.[3]

Both sets of scales (and the item in each scale correlating highest with total scale score) are shown in Table 1. One of the upper-class scales (2–6) had inadequate reliability and was eliminated from the analysis. For the lower-class scales, reliabilities (KR-20 coefficients) ranged from .59–.85 (median of .73), and the correlations between illustrative items in Table 1 and their respective scale scores ranged from .72–.92 (median of .83). For the 19 upper-class scales analyzed, reliabilities ranged from .57–.87 (median of .74), and the correlations for illustrative items ranged from .64–.83 (median of .72).

College Experiences. The 1963 survey included questions about experiences often considered critical in shaping the student's motivation to pursue a scientific or scholarly career. These questions concerned the proportion of the student's close friends entering graduate school immediately, exposure to an honors program or to graduate-level courses, exposure to research experiences as an undergraduate, whether he had won certain kinds of honors or awards in college, and his extracurricular activities.

RESULTS

Changes in Plans to Seek Advanced Training

Changes in plans to seek advanced training were studied by comparing the panel's responses to the three surveys. Figure 1 shows that at the beginning of college 52% of the panel members planned to obtain graduate or professional degrees, after 2 years over 61% reported such plans, and at the completion of college 84% expected to pursue such advanced degrees.

---

[3] Complete lists of items in each of the lower-class and upper-class press scales and keys used in scoring responses have been deposited with the American Documentation Institute. A table showing the validities of precollege characteristics in predicting each criterion variable is also included in this document. Order Document No. 8681 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $1.75 for microfilm or $2.50 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

The right-hand portion of Figure 1 shows the percentages of men and women who reported they were enrolling immediately in a graduate or professional school. It is apparent that the aspiration criterion had a lower "difficulty level" than the entry criterion as a test of motivation for advanced training. At the same time there were no men, and only two women, who "passed" the difficult (entry) item while "failing" the easy (aspiration) item.

The disparity between terminal aspirations and behaviors was much greater for women than for men. Thus, 78% of the women said they planned, but only 27% immediately entered, graduate or professional study; in contrast, 89% of the men planned, and 62% immediately entered, such study. Plans to enter elementary or secondary school teaching were probably the most important factor contributing to this sex difference. A large proportion of the women were education majors for whom 1 or 2 year's teaching experience is customary before pursuing a graduate degree. Thus 21% of the women, but only 9% of the men, said they were planning to apply for admission to a graduate or professional school the following year. In response to another question, 44% of the women, but only 23% of the men, said they planned to go to graduate or professional school, "but not this year." [4]

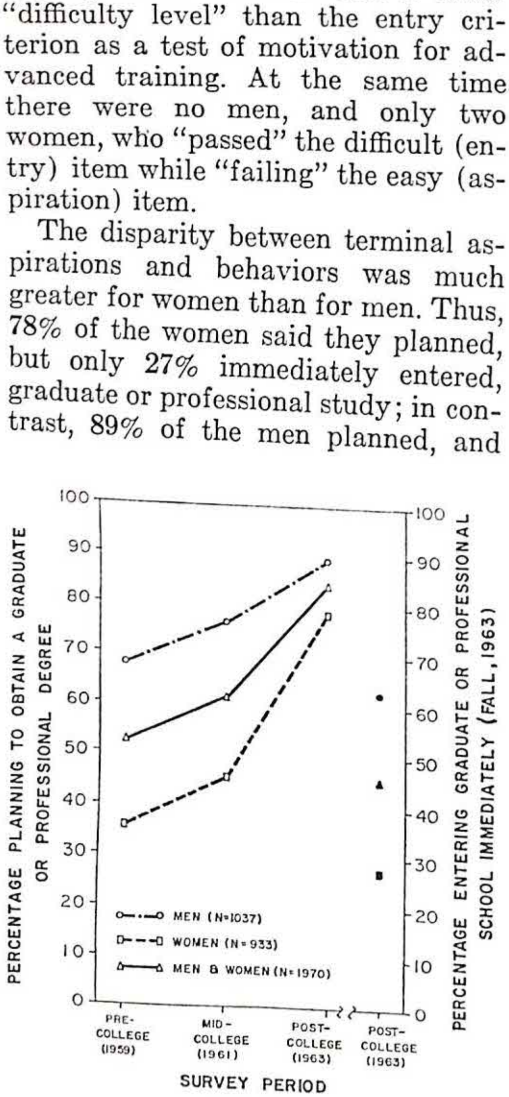


FIG. 1. Changes in plans to seek graduate or professional degrees.

Predictability of Plans

Aspiration Criterion. The total regression analysis yielded a multiple correlation of .507, indicating that, at best, only about 25% of the variance in terminal aspiration levels can be accounted for on the basis of the eight precollege characteristics. Three precollege characteristics—initial aspiration level, sex, and NMSQT score—accounted for virtually all of the predicted criterion variance. Among women there was a multiple correlation of .275 which shrank to .233 when the regression equation for the standardization group was applied to the cross-validation group. Among men the multiple correlation was .451 which shrank to .402 when cross-validated. For both women and men the two predictors accounting for the largest proportions of criterion variance were initial aspiration level and NMSQT score.

---

[4] Academic disqualifications and financial obstacles apparently had little effect upon the student's failure to seek advanced training immediately. Only 2% of all nonentrants applied for admission and failed to receive at least one offer of admission, and only 3% of the nonentrants applied for financial support to attend a graduate or professional school but failed to receive at least one offer of assistance.

Entry Criterion. The entry criterion, like the aspiration criterion, was less successfully predicted for women than for men. For women there was a multiple correlation of .370, which shrinks to .245 when cross-validated, while for men there was a multiple correlation of .418, which shrinks to .332 when cross-validated. The total regression analysis yielded a multiple correlation of .485. In general the best precollege predictors of entry into graduate school tend to be the best predictors of terminal aspirations. The only notable difference was father's educational level: This variable was associated with a significant proportion of entry variance in each of the separate sex analyses, while the unique contribution of this variable to predicting terminal aspirations was negligible.

Optimal and Nonoptimal Residuals

The correlations between optimal and nonoptimal aspiration residuals of cross-validation *S*s were .987 and .973 for men and women, respectively; the correlations between the two kinds of entry residuals were .973 and .970 for men and women, respectively. Clearly the use of the cross-validation technique in the calculation of the criterion residuals would not have materially changed the results. Since the optimal regression equations controlled precollege characteristics most precisely, separate sex analyses are reported below only for the standardization samples.

Factor Structure of College Press Scales

Factor analysis by the method of principal axes, using iterated estimates of communalities in the diagonals and varimax orthogonal rotation, was used to identify the independent dimensions of college press represented in the remaining 33 college press scales. To assess the stability of the emerging factor structures, independent factor analyses were performed for male and female respondents in the standardization and cross-validation samples, respectively. Ten factors were extracted from each of these four analyses. In addition, a similar factor analysis was performed on the intercorrelation matrix based upon scores of both men and women in a pooled sample containing both standardization and cross-validation *S*s. However, only nine factors were extracted in the latter analysis.

Similarities in the factor structures obtained in the five factor analyses can best be indicated by summarizing some of the coefficients of congruence (Harman, 1960, p. 257) calculated for comparing each factor of one analysis with all the factors in another analysis. For example, Factor 1 in the analysis for standardization males had its highest coefficients of congruence with the first factors in the analyses for cross-validation males, standardization females, and for the pooled sample, and parallel statements can be made for Factors 2 through 8. Similarly, the first factor in the analysis for standardization females had its highest coefficient of congruence with the first factors in the analyses for cross-validation women and the pooled sample, and parallel statements may be made for Factors 2 through 8. When analogous coefficients of congruence were computed for the ninth factor, Factor 9 in one analysis had its highest degree of similarity with Factor 9 in another analysis in four out of five of the comparisons. The exception: Factor 9 for standardization males corresponded most closely to Factor 1 for standardization females. The coefficients of congruence for the matching factors ranged from .73–.99 (median of .975). The high degree of consistency and the high coefficients of congruence indicate that the factor structure is extremely stable, particularly for the first seven factors.

It will be recalled that some of the upper-class scales had item content similar to lower-class scales. The following similar pairs of scales were included in the two administrations of press scales: faculty press for compliance (1–5, 2–11), faculty affiliation (1–1, 2–10), faculty enthusiasm (1–3, 2–4), faculty press for achievement (1–4, 2–2), faculty press for independence (1–8, 2–3), student estheticism (1–11, 2–20), and student intellectualism (1–14, 2–13). However, the corresponding scales of each pair typically loaded highly on different factors. This finding was used as a basis for distinguishing between lower-class and upper-class press factors.[5] Also, press scales descriptive of faculty and of students usually loaded highly on different factors, so that most of the factors clearly refer either to faculty press or to student press.

Factor 1 has been called "excellence of faculty in major field." The scales defining this factor, together with their loadings as determined in the "total" factor analysis, are as follows: 2–8: Excellence of faculty teaching (.83); 2–4: Faculty enthusiasm for intellectual values (.81); 2–7: Faculty adequacy as positive role models (.70); 2–12: Faculty structuring of major field of study (.69); 2–3: Faculty press for independent thinking (.68); 2–2: Faculty press for high academic standards (.62); 2–5: Faculty press for advanced training (.54); 2–1: Faculty evaluations of ability (.51); 2–19: Student opposition to faculty influence (−.49).

The loadings suggest a tendency to evaluate teachers in terms of a fairly global "good-bad" dimension. This factor seems similar to the "skill" factor reported by Isaacson, McKeachie, Milholland, Lin, Hofeller, Baerwaldt, and Zinn (1964) who suggest it may represent a "halo effect" in student evaluations. An alternative hypothesis is that this clustering arises from the objects described (i.e., faculty attributes in fact co-vary as students perceive them to co-vary). At present we have no clear evidence for a halo effect in the sense of *spuriously* high correlations between trait attributions. It may be noted that while the upper-class scales for faculty press for enthusiasm, achievement, and independence (2–4, 2–2, 2–3) all load highly on Factor 1, the lower-class counterparts of these scales (1–3, 1–4, 1–8) do not have high loadings on this factor, nor do all the latter scales load highly on the same factor. If one interprets Factor 1 as representing mainly judgmental errors (halo effect), then it becomes necessary to assume that a different pattern of judgmental errors was made in rating teachers on similar items 2 years earlier.

Factor 2 may be called "lower-class humanistic-intellectual press" since the scales defining this factor included only lower-class scales: 1–11: Student press for estheticism (.77); 1–4: Faculty press for achievement (.73); 1–7: Faculty press for humanism (.68); 1–10: Student press for competition (.65); 1–8: Faculty press for independence (.63); 1–12: Student press for reflectiveness (.61); 1–14: Student press for intellectualism (.54). This was the only factor to emerge that was defined by scales clearly descriptive of both faculty and students.

---

[5] It should be noted, however, that neither the items in the foregoing pairs of scales nor the instructions used in administering them were identical. The relatively low correlations may thus have been due to lack of scale equivalence (differences in item content or instructions) or to changes in the kinds of press to which students were exposed during the early and late phases of college.

The third factor, "faculty press for compliance and vocationalism," was the only one defined by scales descriptive of both lower-class and upper-class environments: 1–5: Faculty press for compliance (.76); 1–9: Faculty press for vocationalism (.59); 2–11: Faculty press for compliance (.51). The meaning of Factor 4, "student camaraderie," seems so clear that no argument need be made for its interpretation: 2–17: Student playfulness (.78); 2–16: Student rebelliousness (.58); 2–15: Student cohesiveness (.51); 2–18: Student press for status (.36). Since these scales had no counterparts in the lower-class scales, we cannot interpret this factor as necessarily descriptive of only upper-class environments.

Factor 5 has been called "upper-class student intellectualism," since the lower-class scale on student press for intellectualism had only negligible loadings on this factor: 2–13: Student press for intellectualism (.81); 2–20: Student press for openness to experience (.77).

The sixth factor, "lower-class faculty affiliation," is clearly descriptive of lower-class press, since the upper-class faculty-affiliation scale had only negligible loadings on this factor: 1–1: Faculty press for affiliation (.73); 1–6: Faculty supportiveness (.66); 1–3: Faculty enthusiasm (.58); 1–2: Faculty directiveness (.37).

Factor 7, "upper-class student competition," was defined by a single scale: 2–14: Student press for unfavorable self-evaluations (.43). The lower-class scale on student press for competition had only a negligible loading on this factor.

We are least confident about the interpretation of Factor 8, which has been called "lower-class social conformity." It is best defined by the following scale: 1–13: Student press for social conformity (.35).

In the analyses for cross-validation men and women, Scale 1–13 had loadings of .62 and .69, respectively, on a similar factor. However, in the "total" factor analysis student press for intellectualism (1–14), faculty press for compliance (2–11), and student press for reflectiveness (1–12) had loadings (.32, .31, .30, respectively) almost as high as that for Scale 1–13. It is not clear what these scales have in common.

The ninth factor, "affiliation of faculty in major field," was defined by the following scales: 2–10: Faculty press for affiliation (.77); 2–9: Availability of faculty to students (.76).

The emergence of two independent faculty-affiliation factors (Factors 6, 9) was unexpected. This factor structure emerged much more clearly for men than women, some of whom tended to rate teachers as if the item content of Scales 2–9 and 2–10 was indistinguishable from Factor 1 scales.

College Press and Changes in Plans

To the extent that the correlations between college press variables and the residuals are significantly different from zero we have evidence confirming the hypotheses that the indicated kinds of college press influence dispositions to seek advanced training. We can best avoid overinterpretation of these correlations by confining our attention to those press variables having significant correlations with *both* entry and aspiration residuals. Fourteen of the 33 press scales were significantly correlated with both entry and aspiration residuals of standardization males, while none of the scales were significantly correlated with both kinds of residuals for standardization females. When the correlation between each press factor and the residuals was

TABLE 2

CORRELATIONS BETWEEN COLLEGE PRESS VARIABLES AND RESIDUALS

Press variable	Standardization males ($N = 475$)[a]	
	Entry residual	Aspiration residual
F-1 Excellence of faculty in major field	(.09*)	(.10*)
2–8 Excellence of faculty teaching	.12*	.12**
2–4 Faculty enthusiasm	.10*	.13**
2–7 Faculty adequacy as positive role models	.12**	.17**
2–3 Faculty press for independent thinking	.12*	.14**
2–1 Faculty evaluations of ability	.16**	.14**
2–19 Student opposition to faculty influence	−.13**	−.15**
F-2 Lower-class humanistic-intellectual press		
1–11 Student press for estheticism	.13**	.09*
1–7 Faculty press for humanism	.15**	.10*
1–12 Student press for reflectiveness	.13**	.14**
1–14 Student press for intellectualism	.14**	.11*
F-3 Faculty press for compliance and vocationalism	(−.10*)	(−.11*)
1–9 Faculty press for vocationalism	−.11*	−.11*
F-4 Student camaraderie	(−.10*)	(−.10*)
F-5 Upper-class student intellectualism		
2–13 Student press for intellectualism	.16**	.19**
F-6 Lower-class faculty affiliation		
1–6 Faculty supportiveness	.11*	.12*
1–3 Faculty enthusiasm	.10*	.11*

[a] Factor correlations, in parentheses, are shown only when significant.

* $p < .05$ (two-tailed test).

** $p < .01$ (two-tailed test).

calculated by Dwyer's (1937) method for determining the factor loadings of additional variables, three of the factors were significantly correlated with both residuals for men while none were significantly correlated with both residuals for women. For brevity only the significant press correlations for standardization males are shown in Table 2.

College Experiences and Changes in Plans

There was no sex difference with respect to the number of "experiences" significantly correlated with both entry and aspiration residuals, and Table 3 summarizes only the significant correlations for members of the total sample who responded to each experience item. These results provide support for the hypotheses that the student's disposition to seek advanced training is strengthened by association with peers having high educational aspirations, by teachers who provide favorable evaluations, by undergraduate participation in research, by participation in honors programs, and by winning social recognition for academic achievement.

DISCUSSION

It is difficult to assess the extent to which response biases account for the extraordinary growth of interest in advance training shown by panel members during the college years. However, other survey results suggest that similar, though less marked, increases would be observable in less selective populations of undergraduates who complete college. Thus Davis (1963) found that 77% of the under-

TABLE 3

CORRELATIONS BETWEEN COLLEGE EXPERIENCES AND RESIDUALS

Reported experience	All graduates ($N = 1,654$)[a]	
	Entry residual	Aspiration residual
Peers		
More than half of close friends in college entered graduate or professional school immediately	.18**	.16**
Faculty evaluations		
Undergraduate grade-point average	.20**	.16**
Research experience		
Never participated in original research	−.09**	−.08**
Conducted own research project (e.g., senior thesis)	.10**	.10**
Employed by faculty member as research assistant	.10**	.12**
Participated in summer research training program sponsored by government or foundation	.08**	.10**
Honors program		
Participated in honors program	.07**	.07**
Took one or more graduate-level courses as an undergraduate	.08**	.11**
Awards and honors		
No special honors	−.07**	−.09**
Named to Dean's List	.11**	.08**
Graduated with honors	.08**	.11**
Elected to Phi Beta Kappa	.10**	.08**
Elected to honor society other than Phi Beta Kappa	.11**	.07**
Won scholarship on basis of academic ability	.07**	.05*

[a] *N*s for correlations were less than 1,772 because of nonresponse to specific items concerning experiences.

* $p < .05$ (two-tailed test).

** $p < .01$ (two-tailed test).

graduates he sampled at the end of the senior year of college reported plans to pursue graduate or professional study. Gropper and Fitzpatrick (1959) surveyed undergraduates at the beginning of the senior year and found that 65% planned to seek advanced training; moreover, three-fourths of these seniors reported that they made their decisions to seek, or not to seek, advanced training during the undergraduate years.

Results of the factor analysis of college press scales have implications for future study. We have seen that similar lower-class and upper-class press scales had relatively low correlations and tended to load on different factors. Also press scales descriptive of college faculties and of student peer groups typically loaded highly on different factors. These findings suggest the desirability of defining the object of description in college press scales so that potentially important differences in college subcultures may be more clearly revealed. Also these data suggest that the teacher and peer subcultures to which students are exposed are typically quite different for the lowerclassman and the upperclassman. If it were possible to manipulate college environments so that there was greater constancy in the college press positively related to increased aspirations, then we might expect colleges to have even greater impact upon the educational plans of undergraduates.

It should be noted that the magnitude of the theoretical relationships between college press and changes in degree aspirations cannot be adequately estimated from the correlations in Tables 2 and 3. The observed correlations were undoubtedly reduced by the restricted range of talent of the panel members—response biases and the method of selecting the sample produced a relatively homogeneous group of talented students—and by the fact that no corrections were made for attenuation due to unreliability in the independent variables and particularly in the criterion measures—the use of residual scores entailed considerable criterion unreliability. It seems clear that these correlations would have been larger for a more heterogenous group or if corrections had been made for attenuation.

For the purpose of testing the hypotheses it is sufficient to note whether the correlations are statistically significant. These results strongly suggest that at least some of the types of college press and experiences investigated caused men to change their motivation to seek advanced training. The failure to confirm similar hypotheses for women is congruent with previous studies (Thistlethwaite, 1962a, 1962b).

Finally, brief comment should be made on alternative hypotheses for explaining these correlations. First, one might argue that one or more uncontrolled, precollege characteristics could be correlated with the kinds of teachers and peer groups one will seek out in college, and that these characteristics are reflected in the measures of college environments. Assuming these characteristics are also correlated with the residuals, then one might account for the observed associations through an unmeasured "third" variable. Second, there is the possibility that some of the reported experiences or "measures" of college environments may be indicants of a newly acquired disposition to seek advanced training. Thus once a student resolves to seek advanced training, he may take courses under more demanding teachers or associate with peers who share his career goals. This interpretation simply reverses the direction of causality we have assumed. Third, there is the possibility that the measures of college press share "method" variance with the criterion measures, since both sets of data were obtained from the same *S*s. The last hypothesis is not supported by the results of McFee (1961), Pace (1963), or Saunders (1962). While we may claim that the number of precollege characteristics controlled in the present study exceeds that controlled in any previous study, it was still not possible to rule out all rival hypotheses. This condition does not invalidate the tests of hypotheses; rather it makes these tests less probing than might have been the case under ideal conditions (e.g., if we had available an experimental college in which we could randomly assign *S*s to different student and faculty subcultures). We find it reassuring that a variety of causal hypotheses concerning the effects of college upon the male student's dispositions were exposed to disconfirmation by a fairly searching probe and were not disconfirmed.

REFERENCES

ASTIN, A. W. A re-examination of college productivity. *Journal of Educational Psychology*, 1961, **52**, 173–178.

ASTIN, A. W. Influences on the student's motivation to seek advanced training: Another look. *Journal of Educational Psychology*, 1962, **53**, 303–309. (a)

ASTIN, A. W. "Productivity" of undergraduate institutions. *Science*, 1962, **136**, 129–135. (b)

ASTIN, A. W. Differential college effects on the motivation of talented students *to* obtain the PhD. *Journal of Educational Psychology,* 1963, **54,** 63–71. (a)

ASTIN, A. W. Undergraduate institutions and the production of scientists. *Science,* 1963, **141,** 334–338. (b)

DAVIS, J. A. Great aspirations. Vol. 1. National Opinion Research Center Report No. 90, Chicago, 1963. (Mimeo)

DIRKS, A. W., & ANDERSON, K. E. A study of questionnaire responses from Kansas high school participants and non-participants in the National Merit Scholarship program. *Kansas Studies in Education,* 1960, **10**(2).

DWYER, P. F. The determination of the factor loadings of a given test from the known factor loadings of other tests. *Psychometrika,* 1937, **2,** 173–179.

GROPPER, G. L., & FITZPATRICK, R. *Who goes to graduate school?* Pittsburgh: American Institute for Research, 1959.

HARMAN, H. H. *Modern factor analysis.* Chicago: Univer. Chicago Press, 1960.

HOVLAND, C. I., LUMSDAINE, A. A., & SHEFFIELD, F. D. *Experiments on mass communication.* Princeton: Princeton Univer. Press, 1949.

ISAACSON, R. L., MCKEACHIE, W. J., MILHOLLAND, J. E., LIN, Y. G., HOFELLER, M., BAERWALDT, J. W., & ZINN, K. L. Dimensions of student evaluations of teaching. *Journal of Educational Psychology,* 1964, **55,** 344–351.

KNAPP, R. H., & GOODRICH, H. B. *Origins of American scientists.* Chicago: Univer. Chicago Press, 1952.

MCFEE, ANNE. The relation of students' needs to their perceptions of a college environment. *Journal of Educational Psychology,* 1961, **52,** 25–29.

PACE, C. R. *College and university environment scales: Technical manual.* Princeton: Educational Testing Service, 1963.

SAUNDERS, D. B. *A factor analytic study of the AI and CCI.* Princeton: Educational Testing Service, 1962.

STERN, G. G. *College characteristics index, preliminary manual.* Syracuse: Author, 1958. (Mimeo)

TATSUOKA, M. M., & TIEDEMAN, D. V. Statistics as an aspect of scientific method in research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963. Pp. 142–170.

THISTLETHWAITE, D. L. College environments and the development of talent. *Science,* 1959, **130,** 71–76. (a)

THISTLETHWAITE, D. L. College press and student achievement. *Journal of Educational Psychology,* 1959, **50,** 183–191. (b)

THISTLETHWAITE, D. L. College press and changes in study plans of talented students. *Journal of Educational Psychology,* 1960, **51,** 222–234.

THISTLETHWAITE, D. L. Fields of study and development of motivation to seek advanced training. *Journal of Educational Psychology,* 1962, **53,** 53–64. (a)

THISTLEHWAITE, D. L. Rival hypotheses for explaining the effects of different learning environments. *Journal of Educational Psychology,* 1962, **53,** 310–315. (b)

THISTLETHWAITE, D. L. Diversities in college environments: Implications for student selection. In K. M. Wilson (Ed.), *Research related to college admissions.* Atlanta: Southern Regional Education Board, 1963. Pp. 145–167.

(Received June 7, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 1, 48–53

EVALUATION OF FIVE METHODS OF TEACHING CONCEPTS[1]

DONALD M. JOHNSON AND R. PAUL STRATTON
Michigan State University

Most laboratory studies of concept learning employ a classification method for training and for testing, but teachers employ other methods as well: defining the concept, using it in a sentence, and giving synonyms. The 4 methods and a mixed method were evaluated by comparable training and testing. All single methods were about equally effective; the mixed method was best. Transfer from training by 1 method to tests corresponding to other methods was 100%.

Most concept experiments require the subject (*S*) to learn to classify objects by practice with positive and negative instances and to label the positive instances with a nonsense syllable. Theoretical discussions often take this procedure as typical of concept learning. Carroll (1964) has questioned the relevance of such experiments to concept learning in school because the label is usually a standard word with a network of associations, and because teachers use several methods, some of which are more deductive than inductive. Serra (1953) was able to list a variety of methods by which teachers introduce concepts in elementary school, and inspection of a college textbook will also disclose a diversity of methods. There are at least three, in addition to the classification method, that have been frequently used by teachers and textbooks, but have not been experimentally evaluated: defining a concept, using it in a sentence, and giving synonyms.

Experimental evaluation of these methods requires different tests of concept achievement appropriate to each, and different types of test item have been devised for regular testing of concept achievement in school also. Cronbach (1942) described several types of test item for diagnostic vocabulary testing. Hastings (1941) constructed six tests, each of a different item type, for testing achievement of 35 standard mathematics concepts and attained good reliabilities for each test. These studies have been concerned with blocks of related concepts within an area of knowledge, rather than single concepts, but the results suggest that achievement of single concepts may also be evaluated by different tests which correspond to different teaching methods.

Evaluation of different teaching methods raises the critical question of amount of transfer. Classification practice may build up specific associations that suffice for a classification test, but lack the generality needed for a definition test. However, a recent experiment with 12-year-old children (Johnson & O'Reilly, 1964) found more transfer from classification training to definitions, when the definitions were carefully scored, than previous research had indicated. Since a set to learn concepts facilitates concept achievement (Reed, 1946), college *Ss*, who are used to studying concepts, may show even more transfer. The question can be systematically investigated by giving all tests to all *Ss*.

---

[1] Supported in part by a grant from the National Science Foundation for undergraduate research participation. Thanks are due to Henry Askew for expert assistance in scoring.

This study proposes, therefore, to construct four tests of a few concepts, to teach these concepts by four corresponding methods and by a combination of the four, and to compare achievements on all tests following all teaching methods.

Test Construction and Analysis

Only standard concepts, designated by words, are relevant. The results of a preliminary study with a definition test and a multiple-choice synonym test gave some indication of suitable levels of difficulty. Words with Thorndike-Lorge frequency counts between 11 and 14 could be expected to be easy for college students. Words with counts between 1 and 6 could be expected to be difficult.

Selection was based on frequency and suitability to the four tests to be described. Only nouns, verbs, and adjectives with four synonyms could be used. Words with more than one meaning in an abridged dictionary and technical words that some *S*s might currently be studying were rejected. From the two pools of items that survived editing the experimenters arbitrarily chose four easy words: controversy, vile, rustic, and vicinity, and four difficult words: alacrity, altercation, chide, and opulent.

Tests

There were two tests of the open-end or free-response type and two of the objective or multiple-choice type.

Definitions. Each word was printed with space for a short response. The instructions were: "There are four words to be defined. Explain briefly what each word means. Try each one even if you have to guess."

Sentences. The incomplete-sentence technique is a way of embedding a word in connected discourse. The instructions were: "Here are some incomplete sentences to be finished. Please add a few words to give each sentence a meaningful ending. Try each one even if you have to guess." There were two sentences for each concept, and the eight sentences were printed in irregular order. Two examples follow: "A controversy ceases when____________________." "Something rustic would probably be found ____________________."

Classification. Four short phrases were written to describe objects or events classifiable under each of the four concepts, together with four unrelated phrases. These 20 items appeared in irregular order with the following instructions:

> Most, but not all, of the following items can be classified under four concepts: rustic, vile, controversy, vicinity. Read each item and mark each on your answer sheet thus: rustic 1, vile 2, controversy 3, vicinity 4, none of these 5.

Three sample items follow: "Two men of differing opinions discuss the social value of a book." "A weathered old barn." "An ex-criminal becomes a respected member of the community."

Synonyms. This conventional vocabulary test requires *S* to choose a synonym, but 10 choices were presented rather than the usual 5 in order to increase the reliability of the single item. Four such items were constructed for each concept; the 16 items appeared in blocks of 4, one for each word, in alphabetical order.

The tests were administered in the order given above to reduce the opportunity for learning from the alternatives on the multiple-choice tests. The two free-response tests were printed on separate sheets; each sheet was collected when completed. Time required was about 40 minutes.

Subjects

Group 1, which had the easy concepts, consisted of 63 students in two classes in general psychology at Michigan State University. Group 2, which had the difficult concepts, consisted of 75 students in abnormal psychology. One foreign student was eliminated from each group.

Results

The definitions were rated on a scale of 0–4 by two judges, using ordinary dictionary entries as criteria. The two sentence completions for each concept were treated as a unit and rated on a scale of 0–4 by the two judges because evaluation of a response seemed to be more stable when it was read in conjunction with another response related to the same concept. The criteria of judgment were based on the instructions to write meaningful endings and the assumption that *S*s would try to demonstrate their knowledge. For both easy and difficult concepts, and for both definitions and sentences, the correlations between scorers were satisfactory for present purposes, ranging from .60–.97 with a median of .78. The sums of the two ratings were used in the analysis; hence the scores for each test ranged from 0–8.

The data from Classifications and Synonyms were treated as four-item tests with reliabilities computed by Kuder-Richardson Formula 20. For the easy concepts errors were infrequent, and reliabilities ranged from .19–.72 with a median of .48. For the difficult concepts the range was .65–.89 with a median of .79. The scores were doubled so that the range of scores on each test extended from 0–8.

EVALUATION OF TEACHING METHODS

A comparison of methods requires that each be used separately. Therefore four sets of teaching materials or programs, corresponding to the four tests, were prepared for teaching the four more difficult concepts. A mixed program was also prepared. Mean scores on the four tests of the four difficult concepts ranged from 2.6–5.2 as compared to a maximum of 8, hence there is room for improvement on each.

Materials

Definitions. Definitions were written, with the help of several dictionaries, to characterize each term in a specific way and to place each in a higher-order class. An example follows:

> When two or more people express different opinions, get excited, and contradict each other, the event is called an *altercation.* Thus an *altercation* is a social interaction characterized by heated exchange of opposing arguments. Now write a definition of *altercation* in your own words.

The four definitions, with spaces for the responses, were printed on one side of one sheet of paper; hence all the material remained in view.

Sentences. Russell and Fea (1963) point out that wide reading is universally recommended as a method of building vocabulary, but that direct research on the value of reading in increasing vocabulary has yielded disappointing results. And the reason, they suggest, is the practice of trying to learn unfamiliar words from unfamiliar context.

For present purposes a short story of 174 words was written in which each of the four words appeared twice. The instructions were to read the story and learn the four words. Following Russell and Fea, the context was made easy; to most college students only the four words were unfamiliar. Then *S*s were given four incomplete sentences with the request to finish them, using each of the four words once. All the material remained in view.

Classifications. Short descriptions of objects and events, similar to the test materials, were arranged in blocks of five, one classifiable under each of the four concepts, and one irrelevant. The four words and "none" were printed at the top of each block with instructions to classify each item by writing the appropriate word on a blank line next to the item. One such block appeared on the first page of a booklet, and the next page gave the answers to the first block, followed by a second block of five items. There were six blocks of 30 events to be classified, and correct answers were given to all. The arrangement of the booklet made it easy to turn back and check answers.

Synonyms. The first page of the booklet for the synonym training told *S*s that their task was to learn the meanings of four concepts. Four short statements appeared next: "Alacrity means eagerness." "Altercation means squabble." "Chide means to criticize." "Opulent means luxuriant." Then each of the four concepts appeared next to a blank space, and four other synonyms were given, namely, reproach, quarrel, lavish, and promptness, to be matched with the four concepts by writing one synonym in each blank. The next page gave the answers to these and another block of four synonyms to be matched to the same four concepts. Four blocks of four synonyms were given, followed by answers on succeeding pages. The arrangement of the booklet facilitated checking of answers.

Mixed Program. The review by Serra (1953) reported that teachers commonly use several methods together. Johnson and

O'Reilly (1964) found that children with practice in classifying birds and defining the class learned more than children with only classifying practice. It is reasonable to suppose that, even if the amount of training is equated, a combination of methods would be more efficient than any single method.

The Mixed Program was constructed of materials used in the other four programs with necessary modifications. The definitions were abridged. Each concept appeared only once in the context of a sentence. Two synonyms and one example were given. An illustration follows:

> To chide someone is to talk to him to get him to correct his mistakes. Chide means to criticize or reproach. Thus a mother might chide her children for fighting with each other. An example might be a group of fellows poking fun at a boy with dirty clothes. Now write in your own words what chide means.

This was followed by one block of synonyms for matching and one block of events for classification, with correct answers.

The plan of the experiment required the formulation of certain rules in the preparation of the materials. The four teaching methods should be distinct. The definitions should not include synonyms. The words in context should not include definitions or examples to be classified. In respect to sentences, classifications, and synonyms the items used for teaching and testing should be different. This rule does not apply to the scoring of definitions because memorizing a definition is an accepted part of this teaching method. Each program should be an equally good representative of the method it embodies. This last is a goal that cannot be attained, but the attempt was made to make each program maximally effective.

Each program should require the same time for completion. Differences in reading speed and care make this criterion difficult to attain also, but after some preliminary trials and adjustments it was expected that each program would require about 12 minutes for most college students.

Procedure

Teaching and testing were conducted in a class of about 200 students in general psychology. The five sets of teaching materials, together with an irrelevant set of materials for a control group, were passed out serially so that approximately one-sixth of the class received each set, and neighboring *S*s had different sets. Work was stopped and papers were collected at the end of 12 minutes. Nearly all *S*s had finished.

The tests were given 9 days later. Absence of *S*s on either day reduced the number in the five method groups to 22–28. The control group was increased to 40 by the addition of *S*s who were present on the test day but absent on the teaching day. Scoring of the Definitions and Sentences was carried out by the two judges who had attained the interjudge agreement mentioned above. The sums of the two ratings were used.

Results

Table 1 shows the scores obtained by all groups on the separate tests and also the totals. The maximum score on any test of the four concepts is 32, and the maximum total is 128. In respect to total score all method groups exceeded the control group. These differences are large and, as Table 2 shows, highly significant.

The group that received the Mixed program achieved higher total scores, as expected, than any group that received a single kind of training. These differences are significant with the exception of the group that had training on sentences, and this difference would be significant by a one-tailed test. Within the four groups that received a single kind of training there are no significant differences. For this reason these groups were pooled to form a group of 103 with a mean total of 101.4. The difference between this mean and that of the mixed group is highly significant.

The results for separate tests are similar to those for total scores, but the differences are not so clear-cut. All groups did better than the control group on all tests, and the mixed group did better than any single-method group. On the Definitions test no single-method group is significantly superior to any other, but when the four groups are pooled, the mean of the pooled group is superior to the control group ($p < .01$), and the mixed group

TABLE 1

MEANS AND STANDARD DEVIATIONS OBTAINED ON FOUR TESTS OF CONCEPT ACHIEVEMENT BY SIX GROUPS WITH DIFFERENT TRAINING METHODS

Method groups	N	Achievement scores				
		Def.	Sen.	Class.	Syn.	Total
Definitions	28	20.2 8.4	23.0 8.3	28.0 5.3	27.9 5.3	99.1 22.6
Sentences	26	23.6 7.9	25.2 5.4	28.4 4.0	28.7 4.0	105.8 17.1
Classifications	24	21.4 8.2	24.2 6.6	28.4 5.6	28.8 4.5	102.8 21.0
Synonyms	25	19.9 9.2	24.3 6.8	26.1 4.4	27.6 4.6	97.9 22.1
4 groups pooled	103	21.3 8.5	24.1 6.2	27.7 4.4	28.3 4.6	101.4 20.8
Mixed program	22	26.5 6.5	27.4 5.3	29.3 3.2	30.5 2.0	113.6 11.4
Control	40	12.8 9.9	17.6 8.3	23.5 8.0	23.4 9.0	77.2 31.3

is superior to the pooled group ($p < .01$). The same statements can be made for the Sentences test and the Synonyms test. On the Classifications test the difference between the pooled-group mean and the control-group mean is significant ($p < .01$), and the difference between the mixed-group and the pooled-group mean yields a t of 1.56. In general it is safe to say that each kind of training was effective in raising scores on all tests above the control group and that the mixed training was best.

Another way to look at the data is to examine the mean scores on each test of each concept. Each of the 16 concept means for the pooled group exceeded the corresponding mean for the control group, and 15 of the 16 means for the mixed group exceeded the corresponding means for the pooled group.

TABLE 2

SIGNIFICANCE OF DIFFERENCES BETWEEN TOTALS SHOWN IN TABLE 1

Method groups	Sen.	Class.	Syn.	Mixed	Control
Definitions	1.23	.61	.20	2.95**	3.34**
Sentences		.55	1.42	1.88	4.75**
Classifications			.80	2.19*	3.89**
Synonyms				3.12**	3.10**
4 groups pooled				2.66**	6.00**
Mixed program					6.56**

* $p < .05$.
** $p < .01$.

DISCUSSION

The results show that construction of short tests of single concepts is feasible. The tests of the more difficult concepts yielded 16 reliable scores in about 40 minutes; each score measures knowledge of a single concept by a different response, and each is sensitive to the improvement resulting from a small amount of practice. The small standard deviations for the mixed group on Classifications and Synonyms suggest a ceiling effect, but this effect was not serious enough to prevent this group from exceeding the others.

Concepts of this kind can be taught to college students equally well by different programs of instruction which require different responses. The superiority of the Mixed program supports the common practice of teachers and textbook writers.

The question about transfer may be phrased in this form: How much of what was learned by one method transferred to the tests constructed of materials used in the other methods? Since transfer effects are usually small, one might assume that each of the four single-method groups would get its highest score on the test corresponding to that method. The results of Table 1 do not support this assumption. The group that had practice in writing definitions actually was below the pooled-group mean on the Definitions test, but above the pooled-group mean on Classifications. The synonyms group was the lowest of the four on Synonyms; it was above the pooled-group mean only on Sentences. Overall, there is no evidence of specific relations between training method and test achievement; the picture is one of 100% transfer.

One interpretation for the large transfer effect is that *S* sets concept acquisition as his goal, varying his methods on his own initiative and testing himself so that the intended differences between training methods vanish. Thus each method takes over some of the advantages of the mixed method. Another possibility is that the transfer occurs on the testing day when *S* treats the four tests as problems and uses whatever he can recall to solve them. One with training on sentences may recall the story, for example, and use this information to formulate a definition or to choose a synonym. Probably transfer occurs during both learning and testing.

The matter of interference should be mentioned. One might suppose that if *S* is given a definition or a synonym, there would be no more for him to learn. But he is given four definitions, or four synonyms, as well as verbal instructions; hence some interference can be expected. This may explain why the Sentences method, which depends on learning from context, usually considered inefficient, was as good as any. The continuity of the story probably reduced the amount of interference.

REFERENCES

CARROLL, J. B. Words, meanings and concepts. *Harvard Educational Review*, 1964, **34**, 178–202.

CRONBACH, L. J. Analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research*, 1942, **36**, 206–217.

HASTINGS, J. T. Testing junior high school mathematics concepts. *School Review*, 1941, **49**, 766–776.

JOHNSON, D. M., & O'REILLY, C. A. Concept attainment in children: Classifying and defining. *Journal of Educational Psychology*, 1964, **55**, 71–74.

REED, H. B. Factors influencing the learning and retention of concepts: I. The influence of set. *Journal of Experimental Psychology*, 1946, **36**, 71–87.

RUSSELL, D. H., & FEA, H. R. Research on teaching reading. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963. Pp. 865–928.

SERRA, M. C. How to develop concepts and their verbal representation. *Elementary School Journal*, 1953, **53**, 275–285.

(Received June 17, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 1, 54–60

IMPORTANCE OF PROBLEM HETEROGENEITY TO PROGRAMED INSTRUCTION[1]

ROSS E. TRAUB
McGill University

In this study, different kinds of problems were used to provide experience with the subtasks of a more complex task. The task was graphical integer addition. It was taught by a learning program to 294 6th-grade children who were divided into 3 groups matched for arithmetic ability. The program of 1 group contained 20 heterogeneous subtask problems. Another group got 20 homogeneous problems, and a control group got 20 irrelevant problems. After finishing the program, all students worked 26 graphical integer-addition problems. These were performed best by the heterogeneous-problems group, and this result was found to be independent of student aptitude. Other results suggest that the heterogeneous subtask problems were better because they reduced the probability of making stereotyped or omitted-response errors.

In developing self-instructional programs for relatively complex tasks, such as solving linear equations in one unknown or adding signed numbers, a programmer faces the problem of identifying the skills needed to perform the task. Guidance in dealing with this problem may be obtained from the conceptual framework suggested by Gagné (1962). This framework consists of the analysis of a task into a set of subtasks, all logically necessary for the performance of the task. From this point of view, the skills needed to perform the task are those required for working the subtasks. Programing then involves devising ways to provide experience with the subtasks, the object being to develop the separate subtask skills and promote transfer of those skills to the performance of the task as a whole.

Experience with subtasks may be provided in at least two ways. One is to use instructions that tell about the subtasks (Gagné, 1962). An additional device, commonly used in teaching, is to give practice solving problems of the subtasks.

Several studies have used problems to provide subtask experience (Gagné, Mayor, Garstens, & Paradise, 1962; Traub, 1963; Tuckman, 1962). In general, these studies show that experience obtained from instructions *and* solving problems results in better performance of the task as a whole than experience obtained from instructions alone. However, the *kind* of problems used to provide subtask experience has not been systematically studied. In particular it is not known whether problem heterogeneity, that is, variation in the contexts and the answers of a set of problems, is a factor in learning.

The main purpose of this experiment was to study problem heterogeneity as a factor in learning a complex task from a self-instructional program. To do this, a task was analyzed into subtasks. On some of the subtasks, one group of subjects was given heterogeneous problems to solve. These problems were heterogeneous in that they differed widely in their contexts and in several aspects of their solutions. Another group solved homogeneous problems, ones that were very similar in their contexts and in their solutions. A control group worked problems that reviewed the learning done on previous subtasks. This group was to provide a base line with which to compare the other two groups. After learning the subtasks, all subjects were tested on problems of the task as a whole. Performance of these problems was used to evaluate the effects produced by the different kinds of subtask problems.

---

[1] This study is based on a dissertation submitted to the faculty of Princeton University in partial fulfillment of the requirements for the PhD degree. The research was conducted while the author was an Educational Testing Service Psychometric Fellow. Financial support for the study was obtained from Educational Testing Service and Office of Naval Research Contracts Nonr-1858 (15), Harold Gulliksen, principal investigator, and Nonr-2214 (00), John W. French, principal investigator.

Another aspect of the present experiment was to investigate the relation between problem heterogeneity and subject aptitude. The question asked was whether heterogeneous problems are better at one ability level and homogeneous problems at another ability level. To answer this question, scores on several aptitude and achievement examinations were used to test for an interaction between level of subject aptitude and the effects produced by the different kinds of subtask problems.

Method

The Learning Task and the Experimental Variable

The task used in this study was graphical addition of positive and negative integers. Graphical addition involves drawing an arrow on the number line to demonstrate the addition of two numbers. This is illustrated in Figure 1. Notice that the arrow starts at the location on the number line of the first integer to be added. The length and direction of the arrow represent the second integer so that the stopping point of the arrow locates the sum of the two numbers.

The graphical addition task was analyzed into subtasks, and a program of instructions was written by following the suggestions of Gagné (1962, p. 357). The program had four parts. The first part taught how to locate positive integers on the number line and how to draw arrows to represent positive integers. Part 2 of the program provided similar training for negative integers.

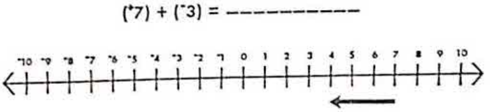


Fig. 1. A solved graphical integer-addition problem.

The third part of the program taught how to draw an arrow of specified length from the number-line location of a specified integer. This training involved two subtasks, one for arrows drawn in a positive direction, the other for arrows drawn in a negative direction. These two subtasks included most of the elements of the graphical addition task, but they did not involve addition problems like the one presented in Figure 1. Instead, these subtasks taught how to draw arrows in response to explicit instructions. In effect the instructions told the student to place his pencil at a certain point on the number line and draw the arrow representing a given integer. Then the student completed the task by identifying the integer found at the end of the arrow.

Three different kinds of problems were devised for the subtasks included in the third part of the program. These different problems constituted the experimental variable and may be described as follows.

Heterogeneous Problems. One set of 20 problems was made heterogeneous in two respects. First, the contexts of the problems were varied by providing a different 20-unit segment of the number line with each problem. Second, the heterogeneous problems were constructed so that their answers varied in certain ways from one problem to the next. Specifically, the integer at the start of an arrow drawn to answer a heterogeneous problem could be larger or smaller in absolute size than the integer at the end of the arrow. Also, the integer at the start of an arrow could have the same algebraic sign as the integer at the end of the arrow, or the signs could be different. These properties of heterogeneous problems are illustrated by the number-line segments and the arrows in Part A of Figure 2.

Homogeneous Problems. A second set of 20 problems was homogeneous. The contexts of these problems were held constant by providing the same segment of the number

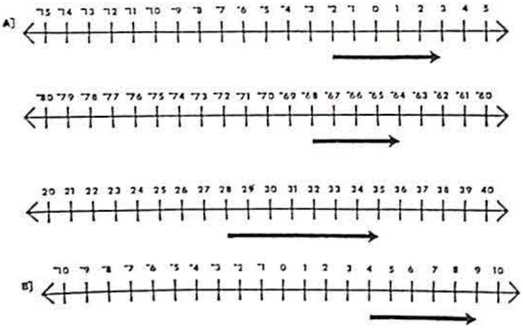


FIG. 2. Number-line segments and correct arrow responses for three heterogeneous subtask problems (Part A) and one homogeneous subtask problem (Part B).

line with each problem. Also the answers to the homogeneous problems were very similar. The integer at the start of the arrow drawn for a homogeneous problem was always smaller in absolute size than the integer at the end of the arrow, and the starting and ending integers always had the same algebraic sign. The number-line segment and the arrow of a homogeneous problem are illustrated in Part B of Figure 2.

Review Problems. A third set of 20 problems was not directly related to the skills taught in the third part of the program. These problems gave practice in drawing arrows to represent integers. That is, these were review problems giving additional practice of the skills learned in the first two parts of the program.

The fourth part of the learning program taught the graphical addition task and presented 26 graphical addition problems. These problems were designed to cover a wide range of difficulty. They consisted of one- and two-digit integers presented for addition in the format illustrated in Figure 1. Each problem required two answers: the graphical one of an arrow drawn on the number line, and the integer answer.

Subjects

The subjects were 294 sixth-grade children who were naïve about graphical integer addition. Of the total sample, 167 (93 boys) came from a public school district in Princeton, New Jersey. The remaining 127 students (76 boys) came from another public school district located near Princeton. All but 57 of the sixth-grade students of the two school districts were included in the sample. The students who were excluded were absent from school for 1 or more days of the experiment.

Procedure

The learning program was presented in the form of a 150-page booklet. To work the program, each student read a page of the booklet, made a written response, and then checked his answer by looking on the back of the page. Wrong answers were corrected before the student continued. This was the procedure for the first 130 pages of the program. The last 20 pages were for the 26 graphical addition problems. Students could not check their answers to these problems.

As was previously indicated, the program had four parts. It was administered on 4 successive days, one part being given each day. On the first day, instructions for working the program were read. These instructions were similar to those used by Gagné et al. (1962). Then the students worked Part 1 of the program. When each student was finished, he recorded a number indicating how long he had taken to complete the first part. (This was also done for the other parts of the program as they were completed.) On the second day of the experiment, Part 2 of the program was administered. Parts 1 and 2 were the same for all students.

For the third part of the program, worked on Day 3, the total sample of children was divided into four sections on the basis of their sex and school district. Each of these four sections was further divided at random so that the students in each of three subsections had the same range of arithmetic aptitude as measured by the Arithmetic Concepts Test of the Iowa Tests of Basic Skills. (The children were first divided by sex and school district because of pronounced differences in the distributions of scores on the Arithmetic Concepts Test. This test was used to equate the sample subsections for arithmetic aptitude because in a pilot experiment a correlation of .74 was obtained between the Arithmetic Concepts Test and a test of integer addition.)

It will now be useful to talk in terms of three experimental groups constructed as follows: Each of the four sections formed by dividing the students on the basis of sex and school district contributed one of its three subsections to each of the three experimental groups. Thus an experimental group was made up of students of each sex from both school districts.

The experimental groups were given the same program of instructions on Day 3, but each got different subtask problems to work.

One group got the 20 heterogeneous problems, another the 20 homogeneous problems, and the third the 20 review problems.

The fourth part of the program was worked on the fourth day. It was the same for all the students and contained the 26 graphical integer-addition problems. Responses to these problems were scored for the number of correct arrow answers and the number of correct integer answers. Also, the arrow errors were qualitatively analyzed and categorized.

RESULTS

The Effect of Problem Heterogeneity on Graphical Integer Addition

The main results of this study are presented in Table 1 which gives the means and standard deviations for two measures of the performance of the experimental groups on the graphical integer-addition problems. It is clear that the best performance was achieved by the heterogeneous-problems group: Both the arrow and integer means of this group were larger than the corresponding means of the homogeneous- and review-problems groups. Also evident from Table 1 is the fact that the homogeneous-problems group was not consistently superior to the review-problems group: The former had a larger arrow mean while the latter had a larger integer mean.

These results were analyzed by one-way analysis of variance. Differences among the experimental groups were significant for both the arrow-response measure ($p < .01$) and the integer-response measure ($p < .05$). Additional analyses using the Newman-Keuls procedure for group comparisons (Winer, 1962, p. 80) demonstrated that the arrow and integer means of the heterogeneous-problems group were significantly greater (p's $< .05$) than the corresponding means of the other two groups. However, the arrow and integer means of the homogeneous-problems group and the review-problems group were not reliably different.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF ARROW AND INTEGER SCORES FOR EACH EXPERIMENTAL GROUP

Experimental group	*N*	Arrow mean	*SD*	Integer mean	*SD*
Heterogeneous problems	97	11.98	8.78	14.81	7.42
Homogeneous problems	93	8.85	7.80	11.90	7.04
Review problems	104	7.11	8.34	12.30	7.86

The Interaction between Problem Heterogeneity and Subject Aptitude

No evidence was found of a reliable interaction between the effects of problem heterogeneity on graphical integer addition and subject aptitude. The test for an interaction was made in two ways. First a two-way analysis of variance design was used in which the different experimental groups provided the categories for one main effect, and several levels of arithmetic aptitude as measured by the Arithmetic Concepts Test provided the categories for the second main effect (see the Treatments × Levels design of Lindquist, 1953, pp. 127–132). Eight such analyses were carried out, one for the arrow scores and one for the integer scores from each of four sections of the total sample. The four sections were obtained by dividing the subjects on the basis of sex and school district.[2] Only one of the eight analyses of variance had a significant interaction term (p

---

[2] Analysis of variance was performed separately on the data from each of the four sample sections because the sections contained unequal numbers of subjects and covered different levels of arithmetic ability. These conditions made it impossible to perform a higher-way analysis of variance including sex and school district as main effects.

< .05). It was interpreted as being caused by chance factors.

The second test for an interaction employed analysis of covariance (Gulliksen & Wilks, 1950). Two analyses were made, one with the arrow scores, the other with the integer scores as the dependent variable. For both these analyses, the independent variables were scores on several aptitude and achievement tests that were administered to the students as part of a more detailed study (Traub, 1964). The results confirmed the conclusion derived from the two-way analyses of variance: The heterogeneous subtask problems produced significantly better performance of the graphical integer-addition task independently of subject aptitude.

Other Results

Classes of Arrow-Response Errors. Three broad classes of errors were identified in the arrow responses to the graphical integer-addition problems. These were stereotyped, random, and omitted-response errors. Stereotyped errors had certain well-defined properties in relation to the problems they were to answer. For example, one stereotyped error had the arrow starting correctly at the location on the number line of the first of two integers being added. Instead of representing the second integer, these arrows were drawn *to the location* of the second integer. There were six other clearly identifiable kinds of stereotyped errors. The second class of errors was random; they could not be related in any way to the problem being answered. The third error type was the omitted response in which no arrow was drawn. Table 2 gives the proportions of arrow answers classified in the three error categories for each experimental group.

TABLE 2

Proportion of Arrows Classified as Stereotyped, Random, and Omitted-Response Errors for Each Experimental Group

Experimental group	Error category			Total
	Stereotyped	Random	Omitted-response	
Heterogeneous problems	.210	.221	.108	.539
Homogeneous problems	.253	.164	.242	.659
Review problems	.388	.185	.156	.729

The most interesting result in Table 2 involves the heterogeneous-problems group. This group made the lowest proportion of stereotyped and omitted-response errors, and the highest proportion of random errors.

The Effect of Heterogeneity on the Difficulty of Subtask Problems. Table 3 presents means and standard deviations for each experimental group on two measures of performance on Part 3 of the learning program. These measures were the number of errors made in working the 20 subtask problems and the time required for Part 3 of the program. It is clear from Table 3 that the heterogeneous-problems group committed more errors and required a longer working time than the other two

TABLE 3

Means and Standard Deviations for Each Experimental Group on Two Measures of Performance on Part 3 of the Learning Program

Experimental group	Errors on subtask problems		Working time (in min.)	
	M	*SD*	*M*	*SD*
Heterogeneous problems	3.43	3.94	34.53	8.78
Homogeneous problems	2.13	3.52	29.35	8.34
Review problems	.10	.32	24.22	6.39

experimental groups. The homogeneous-problems group had the second highest mean score on both measures, while the review-problems group had the lowest. All differences between the group means reported in Table 3 were significant (p's $< .01$) when tested by the Newman-Keuls procedure. These significant group differences contrast with the nonsignificant differences found among the time and error means for Parts 1 and 2 of the program. Thus it follows that the differences obtained for Part 3 of the program were produced by differences in the difficulty of the three kinds of subtask problems. The heterogeneous problems were most difficult, the homogeneous problems second most difficult, and the review problems easiest.

DISCUSSION

In this study, students who worked heterogeneous subtask problems performed the task significantly better than students who worked either the homogeneous problems or the irrelevant review problems. One reason for this finding is suggested by the results of the error analysis of responses to the graphical integer-addition problems. The experimental groups differed in the proportion of arrow responses that could be classified as stereotyped errors. In particular, the heterogeneous-problems group made the smallest proportion of stereotyped errors. Also, this group omitted the smallest proportion of responses. Thus, the heterogeneous-problems group made fewer stereotyped errors while answering more problems.

It is clear that one reason the heterogeneous-problems group did better on the final task was because they made fewer stereotyped errors and omitted fewer responses. The reason for this is probably that the heterogeneous problems contained more information about the task than the other kinds of subtask problems. Support for this suggestion is provided by the finding that the heterogeneous problems were the most difficult. It is likely that difficulty has a high correlation with the amount of information contained in a set of problems.

Additional and logically more direct support for the information interpretation comes from an inspection of the heterogeneous subtask problems and the graphical integer-addition problems. The fact is that both these sets of problems were heterogeneous in the sense of presenting problems in the context of several different number-line segments. Moreover, the different kinds of arrow answers required for the heterogeneous subtask problems were also required for the graphical integer-addition problems.

From the foregoing it is clear that, while all groups had the same instructions for solving the graphical integer-addition task, the heterogeneous-problems group had more information about certain aspects of the task. The homogeneous- and review-problems groups did not have the information, and hence for these groups the graphical integer-addition problems would have been relatively more complex. Apparently the effect of this greater complexity was to increase the probability of forgetting or confusing the correct procedure for solving the task. Consequently, the homogeneous- and review-problems groups adopted a problem-solving procedure that resulted in a stereotyped error, or else they gave no response at all.

In summary, the findings of the present investigation indicate that heterogeneous subtask problems can produce better learning of a complex task such as graphical integer addition when the teaching is done by a self-instructional program. The importance of heterogeneous problems is additionally emphasized by the negative finding of the study. This finding was that the different kinds of subtask problems affected learning independently of level of subject aptitude. Heterogeneous problems were best for students of high and low aptitude.

REFERENCES

Gagné, R. M. The acquisition of knowledge. *Psychological Review*, 1962, **69**, 355–365.

Gagné, R. M., Mayor, J. R., Garstens, Helen L., & Paradise, N. E. Factors in acquiring knowledge of a mathematical task. *Psychological Monographs*, 1962, **76** (7, Whole No. 526).

Gulliksen, H., & Wilks, S. S. Regression tests for several samples. *Psychometrika*, 1950, **15**, 91–114.

Lindquist, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

Traub, R. E. Basic factor abilities: Their relation to rate of learning and to the effect of repetition on performance of a programmed learning task. Unpublished minor research paper, Princeton University, 1963.

Traub, R. E. The importance of problem heterogeneity to programmed learning. Princeton, N.J.: Educational Testing Service Research Bulletin No. RB-64-26, Office of Naval Research Technical Report, May 1964.

Tuckman, B. W. Ability and practice in the performance of non-reproductive tasks. Unpublished minor research paper, Princeton University, 1962.

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received October 15, 1965)

Valuable McGraw-Hill Books

PSYCHOLOGY OF LEARNING AND TEACHING, Second Edition

By HAROLD W. BERNARD, Oregon State System of Higher Education. 500 pages, $7.95.

The major revision shows how intelligence, motivation, home, peers, teachers, and content of subjects shape the desire for and efficiency, economy, and permanence of learning.

DISCIPLINE AND CHILD GUIDANCE

By KARL S. BERNHARDT, University of Toronto, with Foreward by W. E. BLATZ. *McGraw-Hill Series in Guidance, Counseling, and Student Personnel in Education.* 322 pages, $6.50.

Formulates a sound and functional philosophy of child rearing, based on the author's extensive experience with The Institute of Child Study at the University of Toronto.

CHILD DEVELOPMENT, Fourth Edition

By ELIZABETH B. HURLOCK, University of Pennsylvania. McGraw-Hill Series in Psychology. 730 pages, $7.95.

Covers the entire scope of the child's development from conception to adolescence.

EDUCATIONAL PSYCHOLOGY: The Principles of Learning in Teaching

By ARDEN N. FRANDSEN, Utah State University. *Psychology and Human Development in Education.* 640 pages, $7.50.

The purpose of this text is to aid teachers in meeting the challenge of modern education: the provision of *better* education for children and youth.

EDUCATIONAL PSYCHOLOGY

By WALTER B. KOLESNIK, University of Detroit. *McGraw-Hill Catholic Series in Education.* 560 pages, $7.95.

This book is intended for use by prospective elementary and high school teachers at the undergraduate level in Catholic colleges and universities.

INTRODUCTION TO THE GIFTED

By GERTRUDE HILDRETH, Professor Emeritus, Brooklyn College, The City University of New York. Psychology and Human Development in Education. 576 pages, $7.95.

Reflects the author's personal contact with gifted children over a period of 40 years.

THE MENTALLY RETARDED CHILD: A Psychological Approach

By HALBERT B. ROBINSON and NANCY M. ROBINSON, both of the University of North Carolina. *McGraw-Hill Series in Psychology.* 640 pages, $10.50.

The only comprehensive, up-to-date text for psychologists, educators, and other workers entering the field of mental retardation.

EDUCATING EMOTIONALLY DISTURBED CHILDREN

By NORRIS G. HARING, University of Kansas Medical School; and E. LAKIN PHILLIPS, National Orthopedic and Rehabilitation Hospital, Arlington, Virginia. *Psychology and Human Development in Education.* 371 pages, $6.95.

An upper division and graduate level text stressing the simplicity and practicality of *a structured program* for emotionally disturbed children in a school setting.

PSYCHOLOGY IN EDUCATION, Fourth Edition

By HERBERT SORENSON, University of Kentucky, 555 pages, $7.95.

This introductory text provides clear understanding and guidance in practical applications of psychological theory.

UNDERSTANDING EDUCATIONAL RESEARCH, Second Edition

By D. B. VAN DALEN, University of California, Berkeley; and WILLIAM J. NEYER, Syracuse University. Psychology and Human Development in Education. Available in April.

Reflecting recent changes in educational research, this edition introduces the novice to the tactics and terminology of research work in clear, easy-to-understand language.

SEND FOR YOUR EXAMINATION COPIES TODAY

McGraw-Hill Book Company

330 WEST 42ND STREET / NEW YORK, N. Y. 10036

Forthcoming

1966

An Introduction to Statistics and Measurement:

A Programmed Book by Lowell A. Schoer, State University of Iowa. Concisely presents statistical concepts. Thoroughly class-tested Little mathematical background needed. Paperbound $4.95 tent.

Career Information in Counseling and Teaching

by Lee E. Isaacson, Purdue University. Stresses preparation and entrance to the world of work at all occupational levels and the psychological and sociological factors which bear upon the worker and his job. $8.50 tent.

The Causes of Behavior:

Readings in Child Development and Educational Psychology, Second Edition by Judy Rosenblith, Wheaton College and Wesley Allinsmith, University of Cincinnati. Second Edition completely up-to-date. Latest research studies.

Human Development in Western Culture,

Second Edition by Harold W. Bernard, Oregon State System of Higher Education. Organized around the concept of developmental tasks. Stresses the role of social class in development. Examines the current impact of the Economic Opportunity Act and the Elementary and Secondary School Act of 1965.

ALLYN AND BACON, INC.
Dept. Z1 150 Tremont Street, Boston, Mass. 02111

Have you seen these?

Teaching the Gifted Child

by James J. Gallagher, University of Illinois. Delineates the differences between and characteristics of high I.Q. children, highly creative children, and gifted underachievers. Demonstrates the importance of changes in content, with chapters devoted to science, mathematics, and social studies. $7.50

Teaching Gifted Students:

A Book of Readings by James J. Gallagher, University of Illinois. Follows the organizational structure of *Teaching the Gifted Child* by the same author but may be used in conjunction with any basic text. Paperbound $4.95

Evaluating Pupil Growth,

Second Edition by J. Stanley Ahmann, Colorado State University and Marvin D. Glock, Cornell University. A completely up-to-date presentation of the construction and use of various tests in all aspects of formal education. $7.95

Educational Psychology,

Second Edition by James M. Sawrey and Charles W. Telford, both of San Jose State College. New chapter on the nature and scientific methods of educational psychology. Expanded treatment of measurement. $7.95

ALLYN AND BACON, INC.
Dept. Z1 150 Tremont Street, Boston, Mass. 02111

New books of interest

SECONDARY REINFORCEMENT

Selected Experiments

EDWARD L. WIKE This is the first full integration of present knowledge regarding positive secondary reinforcement. The text isolates functional relationships and exemplifies them by presenting an organized set of experimental studies in their complete, original form. The Introduction and Resume consider such topics as Hull's Approach to Motivation, Skinner's Approach to Motivation, Paradigms for Secondary Reinforcement, and Principles of Secondary Reinforcement. 516 pp. $13.75

A PRACTICAL INTRODUCTION TO MEASUREMENT AND EVALUATION

Second Edition

H. H. REMMERS, N. L. GAGE & J. FRANCIS RUMMEL Retaining the book's original organization, the Second Edition focuses on the theories and techniques of statistical measurement; the administration, scoring, and interpretation of tests; the evaluation of classroom instruction through teacher-made tests; and the appraisal of emotional and social adjustment. There is a new section, "Taxonomy of Educational Objectives: Affective Domain," and discussions of the status of the measurement of creativity and of *Digitek*, the new scoring machine. 390 pp. $6.95

Harper & Row Publishers

49 East 33d St., New York 10016

INTERVIEWERS
MARKET RESEARCH
PRIMARILY IN THE MEDICAL FIELD

* * *

Part-time work, assignments adjustable to available time: Ph.D. preferred, M.A. considered: Interviewing experience required, but medical experience not essential: Any location in the United States. High per interview rate plus bonus.

* * *

Address inquiries to:

THE MENDOTA
RESEARCH GROUP
31 SYLVAN AVENUE
ENGLEWOOD CLIFFS,
NEW JERSEY

What are psychologists doing about

700,000 school dropouts annually?

Find out in these eight new Wiley books . . .

GUIDANCE in TODAY'S SCHOOLS

Second Edition

By DONALD G. MORTENSEN, *California State College;* ALLEN M. SCHMULLER, *Purdue University.* Used by over 100 colleges and universities and now thoroughly updated in its second edition, this timely book explores the possibilities of guidance in the school as a means to helping each student to realize his full potential. The authors emphasize four essential phases of the guidance program: understanding of each individual by self and others; knowledge of social needs and values; programs for healthy adjustments; and provisions for challenging educational experiences. *1966. Approx. 480 pages. $6.95.*

WORKING with GROUPS

Group Process and Individual Growth

Second Edition

By WALTER M. LIFTON, *Coordinator of Pupil Personnel Services, City School District of Rochester, New York.* Widely praised in its first edition, this book shows how the proper use of groups can free the individual and allow him to become a fruitful resource in a society faced with shortages of manpower in the helping professions. *1966. Approx. 304 pages. $6.95.*

SOCIALIZATION AFTER CHILDHOOD:

Two Essays

By ORVILLE G. BRIM, JR. *and* STANTON WHEELER, *both of the Russell Sage Foundation.* Provides an orienting perspective on problems posed by the study of socialization and examines the large bureaucratic organization, an increasingly important setting in which post-childhood socialization occurs. *1966. Approx. 144 pages. Paper: Prob. $2.95. Cloth: Prob. $5.95.*

STUDYING the CHILD in SCHOOL

By IRA J. GORDON, *University of Florida. 1966. In Press.*

MENTAL HEALTH and ACHIEVEMENT

Increasing Potential and Reducing School Dropout

Edited by E. PAUL TORRANCE, *University of Minnesota; and* ROBERT D. STROM, *The Ohio State University.* 28 eminent psychologists, educators, and guidance workers describe how many subtle familial, school, peer, and communal pressures act to impede the mental growth and achievement of today's student. The book not only provides expert analysis on the changing function of the school staff, but it also applies psychological principles to problems teachers must face daily. *1965. 417 pages. $4.95.*

PERSONALITY:

An Objective Approach

By IRWIN G. SARASON, *University of Washington.* A unique treatment that empirically orients various theories. Involving the reader directly in actual research, the book discusses both personality assessment and the experimental study of personality. *1966. Approx. 664 pages. Prob. $7.95.*

INDIVIDUAL DIFFERENCES

Edited by ANNE ANASTASI, *Fordham University.* A thoroughly integrated presentation of selected readings on individual differences, this distinguished collection examines problems of contemporary interest in differential psychology from a historical perspective. *1965. 301 pages. Paper: $2.95. Cloth: $5.00.*

The ADOLESCENT EXPERIENCE

By ELIZABETH DOUVAN *and* JOSEPH ADELSON, *both of The University of Michigan.* The authors have pieced together a composite picture of the American teenager from a mass of information provided by two national interview studies. And the results challenge many popular notions about adolescence. Thus, the book argues that sex differences and a youngster's conceptions of his own future are far more crucial than psychologists have supposed. *1966. 437 pages. Prob. $7.25.*

JOHN WILEY & SONS, Inc. 605 Third Avenue, New York, New York 10016

Macmillan Announces . . . Three New Additions to

The Psychological Foundations of Education Series

A convenient paperbound collection of
specialized studies in education

General Editor: Victor H. Noll,
Michigan State University

The Psychology of Learning In The Classroom

by Robert C. Craig,
Marquette University

A brief, specialized guide designed to be used as supplementary material in conjunction with a comprehensive text in educational psychology or human learning.

1966, 96 pages, $1.25

Problem Solving In The Classroom

by Bryce B. Hudgins,
Washington University

In a practical, non-technical manner, the author presents a variety of research evidence on the processes of problem solving, not as a method of teaching, but as a habit of inquiry that has as its goal the development of new ideas based on old ones.

1966, 96 pages, $1.25

Teacher Self-Evaluation

by Ray H. Simpson, University of Illinois

This concise, informative guidebook offers the most modern procedures for teacher self-evaluation in all key career areas and—of equal importance—presents realistic methods for utilizing reasonable amounts of time and thought to this critical function.

1966, 96 pages, $1.25

Also Available In The Series . . .

Psychology of Adolescence For Teachers

by Glenn Myers Blair and R. Stewart Jones, both of the University of Illinois

1964, 128 pages, $1.50

Psychology Of The Child In The Classroom

by Don C. Charles, Iowa State University

1964, 96 pages, $1.25

The Mentally Retarded Child In The Classroom

by Marion J. Erickson, Ypsilanti Public Schools

1965, 96 pages, $1.25

Guidance In The Classroom

by Ruth Strang, University of Arizona, and Glyn Morris, Assistant Supervisor of Guidance and Curriculum, Lewis County, New York

1964, 128 pages, $1.50

Gifted Children In The Classroom

by E. Paul Torrence, University of Minnesota

1965, 112 pages, $1.25

Write to the Faculty Service Desk for Examination Copies

THE MACMILLAN COMPANY

60 FIFTH AVENUE, NEW YORK 10011

Macmillan Meets Your Teaching Needs With . . . .

Readings In Educational Psychology

Edited by Victor H. Noll and Rachel P. Noll
Michigan State University

Fifty-three compelling readings highlight every important aspect of educational psychology. Teachers tested many of these readings in the classroom; the editors then selected those that captured student interest and attention. Among the many vital areas discussed are: the youngest generation, human relations, mental health, measurement and evaluation, and valuable teaching aids.

1963, 568 pages, paper, $3.95

Educational Psychology, Second Edition

by Glenn Myers Blair, R. Stewart Jones, and Ray H. Simpson, all of the University of Illinois

An unusually complete, comprehensive, and functional treatment of educational psychology with facts, principles, and methods of procedure reinforced with illustrations from the psychology laboratory, clinical experiences, classroom experimentation and related fields. The authors include a separate section devoted to the psychology of the teacher and his role in the learning process. Information is presented on teaching machines, feedback, brainstorming, creative thinking, discovery, and learning.

1962, 678 pages, $7.95

Statistics In Education And Psychology: A First Course

by Merle W. Tate, University of Pennsylvania

This book provides the student with a comprehensive treatment of the basic concepts and procedures that are most useful for measurement and research in education and psychology. It is in part an abridgement and in part a revision of the author's *Statistics in Education*; this text, however, is quite different in purpose and coverage, providing a much more fundamental treatment of statistics. It is intended primarily as a text for a one-semester course for psychology and education students.

1965, 352 pages, $5.95

Write to the Faculty Service Desk for Examination Copies

THE MACMILLAN COMPANY
60 FIFTH AVENUE, NEW YORK 10011

ANNOUNCING A NEW FILM

BEHAVIOR THEORY IN PRACTICE

Ellen P. Reese, *Mount Holyoke College; produced by Charter Oak Studio.* An hour and twenty minutes of 16 mm. color film, with sound and narration, which describes basic research in Skinnerian behavior theory and extends these findings outside the laboratory. These four twenty-minute reels represent a valuable teaching aid in college courses in educational psychology and in introductory, learning, and experimental psychology.

ALSO AVAILABLE

CHILD DEVELOPMENT I: A SYSTEMATIC AND EMPIRICAL STUDY

Sidney W. Bijou and Donald M. Baer, *both of the University of Washington.*
90 pp., paper, $1.25

CHILD DEVELOPMENT II: UNIVERSAL STAGE OF INFANCY

Sidney W. Bijou, *University of Illinois;* and Donald M. Baer, *University of Kansas.*
183 pp., paper, $1.75

EDUCATIONAL PSYCHOLOGY, Second Edition

Karl C. Garrison and Albert J. Kingston, *both of the University of Georgia;* and Arthur S. McDonald, *Marquette University.*
544 pp., illus., $6.50

APPLETON-CENTURY-CROFTS
Division of Meredith Publishing Company
440 Park Avenue South, New York, N. Y. 10016

The 1965 DIRECTORY of the AMERICAN PSYCHOLOGICAL ASSOCIATION

A revised standard reference containing biographical and geographical listings of the Association's 23,500 fellows, members, and associates.

A valuable tool for psychologists and members of related professions
libraries publishers
personnel offices

$10.00 ($5.00 to APA members)

Order Form

American
Psychological
Association
Order Department
1200 17th Street, N.W.
Washington, D.C. 20036

Date..............

Please send.......copies of the 1965 APA Directory to:

....
(please print or type)
....
....

Check enclosed....
Please bill....

Signature ....

announcing the Second Edition

LEARNING AND HUMAN ABILITIES

Educational Psychology

Herbert J. Klausmeier · William Goodwin

Almost completely rewritten and thoroughly updated in the Second Edition, this widely adopted text continues to emphasize the concept of emerging human abilities—thus integrating the notions of growth and learning that are traditionally treated separately. The new content and organization meet a wider range of needs of students and professors. Especially suited to courses which emphasize teaching—learning processes in school settings as well as the development of intellectual abilities the book is ideal for several types of courses: a teaching-oriented course, a science-oriented course, or a comprehensive course which includes measurement and experimental design. Course outlines have been prepared for the instructor, and suggest a variety of combinations of all or parts of each of the 18 chapters. *March.*

A Combined Adjunct Program and Student Workbook is available for use with the text. The Adjunct Program contains 35 to 40 multiple-choice items and about 10 subjective questions per chapter; the Workbook comprises questions that encourage reasoning, evaluation, and creativity. A Student Evaluation Guide is also available.

READINGS IN LEARNING AND HUMAN ABILITIES

Educational Psychology

Richard E. Ripple, Editor

This volume contains forty-nine unabridged readings for use with LEARNING AND HUMAN ABILITIES: *Educational Psychology* or with other introductory educational psychology texts. A cross-reference chart correlates the readings with chapters in sixteen current texts. Although emphasis is on articles reporting empirical research studies, some theoretical articles and survey articles reviewing research are included. A balance is achieved among articles dealing with the elementary, secondary, and college levels. 586 pp. $4.95/paper

HARPER & ROW, PUBLISHERS • 49 E. 33d ST., NEW YORK 10016

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION *(Act of October 23, 1962: Section 4369, Title 39, United States Code)*	Publisher: File two copies of this form with your postmaster.

1. DATE OF FILING 9/29/65

2. TITLE OF PUBLICATION JOURNAL OF EDUCATIONAL PSYCHOLOGY

3. FREQUENCY OF ISSUE Bi-Monthly

4. LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, city, county, state, zip code)*
Waverly Press, Inc., Mount Royal and Guilford Avenues, Baltimore, Maryland 21202

5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*
1200 17th Street, N.W., Washington, D.C. 20036

6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER *(Name and address)*
American Psychological Association, 1200 17th Street, N.W., Washington, D.C. 20036

EDITOR *(Name and address)*
Raymond G. Kuhlen, 123 College Pl., Syracuse 10, New York

MANAGING EDITOR *(Name and address)*
Helen Orr, 1200 17th Street, N.W., Washington, D.C. 20036

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)*

NAME	ADDRESS
American Psychological Association	1200 17th Street, N.W., Washington, D.C. 20036

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

NAME	ADDRESS

9. Paragraphs 7 and 8 include, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting, also the statements in the two paragraphs show the affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner. Names and addresses of individuals who are stockholders of a corporation which itself is a stockholder or holder of bonds, mortgages or other securities of the publishing corporation have been included in paragraphs 7 and 8 when the interests of such individuals are equivalent to 1 percent or more of the total amount of the stock or securities of the publishing corporation.

10. THIS ITEM MUST BE COMPLETED FOR ALL PUBLICATIONS EXCEPT THOSE WHICH DO NOT CARRY ADVERTISING OTHER THAN THE PUBLISHER'S OWN AND WHICH ARE NAMED IN SECTIONS 132.231, 132.232, AND 132.233, POSTAL MANUAL *(Sections 4355a, 4355b, and 4356 of Title 39, United States Code)*

	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	SINGLE ISSUE NEAREST TO FILING DATE
A. TOTAL NO. COPIES PRINTED *(Net Press Run)*	7,500	7,354
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES	none	none
2. MAIL SUBSCRIPTIONS	6,700	6,703
C. TOTAL PAID CIRCULATION	6,700	6,703
D. FREE DISTRIBUTION *(including samples)* BY MAIL, CARRIER OR OTHER MEANS	25	88
E. TOTAL DISTRIBUTION *(Sum of C and D)*	6,725	6,794
F. OFFICE USE, LEFT-OVER, UNACCOUNTED, SPOILED AFTER PRINTING	715	560
G. TOTAL *(Sum of E & F—should equal net press run shown in A)*	7,440	7,354

I certify that the statements made by me above are correct and complete. *(Signature of editor, publisher, business manager, or owner)*

POD Form 3526 Jun. 1965

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 57 April 1966 Number 2

CONTENTS

Group Cohesiveness and Individual Learning 61

ALBERT J. LOTT AND BERNICE E. LOTT

Effect of Instructional Set and Variety of Instances on Children's Learning 74

HARRIETT AMSTER

Reactive Inhibition as Related to Performance on Programed Learning Materials 86

LOWELL SCHOER

Changes in Attitudes and Values Associated with College Attendance 89

IRVIN J. LEHMANN, BIRENDRA K. SINHA, AND RODNEY T. HARTNETT

Overt Versus Covert Responding and Retention by Sixth-Grade Students 99

ROBERT L. CRIST

Sequence of Vocal Response-Differentiation Training and Auditory Stimulus-Discrimination Training in Beginning French 102

LAWRENCE MACE

Blackout Ratio and Overt Responses in Programed Instruction: Resolution of Disparate Results 109

FREDERICK D. KEMP AND JAMES G. HOLLAND

Structure of Experience and Functional Fixedness 115

GEORGE D. YONGE

Journal of Educational Psychology
1966, Vol. 57, No. 2, 61–73

GROUP COHESIVENESS AND INDIVIDUAL LEARNING[1]

ALBERT J. LOTT AND BERNICE E. LOTT
University of Kentucky

Using a learning-theory formulation it was predicted that individuals will perform better on learning tasks if they are in HiCo groups (with highly liked others) rather than in LoCo groups (with less-liked others). Each of 206 4th- and 5th-grade children, classed as high or low in IQ (California Test of Mental Maturity), was placed, on the basis of sociometric information, in a same-sex HiCo or LoCo group. Comparisons, made on 2 verbal learning tasks, retention, and relearning, indicated that high-IQ children (especially girls) in HiCo groups did consistently better than those in LoCo groups. For low-IQ children cohesiveness made no reliable difference although those in LoCo groups tended to do slightly better. Other measures are congruent with the learning-task findings. Results are discussed in terms of drive and task complexity (competing responses).

In an attempt to coordinate small-group behavior to more general behavior principles, the concept of cohesiveness has been formulated in terms of reinforcement learning theory. The present approach assumes that the primary source of a group's attractiveness to its members is the attractiveness of fellow members. The concept of attitude is substituted for that of attraction because the former has a precise and particular meaning within reinforcement learning theory (an anticipatory response with cue and drive properties—Doob, 1947), and cohesiveness is then defined as that group property which is inferred from the number and strength of mutual positive attitudes among the members (Lott, 1961).

The authors have previously tested hypotheses regarding antecedents of cohesiveness (James & Lott, 1964; Lott & Lott, 1960). Theoretically derived predictions regarding the consequences of cohesiveness also follow from placement of the concept within a Hullian framework. In an earlier investigation, degree of group cohesiveness was shown to be positively related to both intragroup level of communication and the tendency of members to conform to a dominant group opinion (Lott & Lott, 1961). The present study was designed to test another predicted consequence of variations in cohesiveness. Since attitudes are presumed to have drive value, the greater the degree of intermember liking (group cohesiveness), the higher should be the general drive level in individual members. Since learning is ordinarily a positive function of drive, the authors predict that better individual learning will take place in high- than in low-cohesive groups (other things being equal), or that individuals who are in the presence of highly liked others will perform better on learning tasks than individuals who are in the presence of less-liked others.

Little previous research has been

---

[1] The research reported herein was supported through the Cooperative Research Program of the Office of Education, United States Department of Health, Education, and Welfare. It could not have been carried out without the cooperation and assistance of four school principals in Lexington, Kentucky—A. Brewer of Ashland School, J. Smith of Constitution School, S. Gabby of Linlee School, and T. Dunn of Mary Todd School—their teachers, and their pupils. We are grateful, also, to May Briscoe and Sherwood Burress for help in various phases of the research.

concerned with learning as a function of interpersonal attraction per se. Perkins (1950) and Rehage (1951) investigated the consequences of different classroom "climates" for learning, or problem solving, but in these studies manipulation of leadership (teaching) techniques constituted the independent variable. More relevant to the present study is the finding by Shaw and Shaw (1962) that early spelling scores were positively correlated with the degree of liking among members of second-grade study groups.[2]

In the present study we compared high- and low-cohesive groups of fourth- and fifth-grade children of similar tested general intelligence on a verbal learning task, retention, relearning, and a second verbal learning task. The authors' expectation that children in high-cohesive groups will learn better and retain more than children in low-cohesive groups presupposes that the children being compared are similar in general learning ability, and it was assumed that the latter is validly reflected by IQ scores. Specifically, the authors predicted that among children who are similar in tested intelligence, either high or low, those who are in small groups with other children whom they like a great deal will learn and retain new material better than those who are with children they like less.

Since it was assumed that objects or persons which evoke positive attitudes can serve as secondary reinforcers (Hull, 1952; Lott & Lott, 1961), that is, can strengthen behavior when they are presented contingent upon that behavior, it was necessary to eliminate as far as possible the chances of this occurring. Such direct intermember reinforcement would confound the effects of increased drive based on the mere presence of liked group members. If the group members were permitted to freely interact, then the children in high-cohesive groups could reinforce a variety of one another's behaviors, related or unrelated to the learning tasks. Consequently, the experimental situations were carefully planned to control and channel member interaction. Both verbal and nonverbal communication among members of all groups was strictly structured and confined to task essentials.

METHOD

Subjects

Four elementary schools, three predominantly white and one entirely Negro in student body, cooperated in this investigation. Within each school, subjects (*S*s) were obtained from one fourth- and one fifth-grade class. For Learning Task I (LTI) a total of 206 children (97 boys, 109 girls) was selected on the basis of sociometric and IQ-score considerations to participate in three- or four-person same-sex groups.

One week after LTI, the same *S*s (reduced in number to 155; 71 boys, 84 girls) were tested, within the same small groups, on Retention (RET), Relearning (REL), and Learning Task II (LTII). *S* losses are accounted for by absences: In the case of three-person groups, the absence of one child eliminated the other two; if one child was absent from a four-person group, however, the group simply became a three-person group for the remaining tasks.

Procedure

Experimental Formation of Groups. Two days prior to presentation of LTI, all members of each of the eight participating classes were given a sociometric test (Socio I) by their regular teacher. This consisted of a Friendship Book in which each child rated all other same-sex children in his class on the following scale:

---

[2] A number of recent investigations have tested hypotheses which relate verbal conditioning to the relationship between *S* and *E*, or simply to the attractiveness of *E* (e.g., Binder, McConnell, & Sjoholm, 1957; Reece & Whitman, 1962; Sapolsky, 1960). Although these studies are theoretically relevant to the present hypothesis, discussion of them will be omitted since the focus here is on learning within a group context where the group is composed of individuals who are mutually attracted to one another to a high or low degree.

________ I like him very, very much.
________ I like him. He's alright, but I like some other boys better.
________ Sometimes I like him a little bit and sometimes I don't like him.
________ Most of the time I don't like him.
________ I never liked him and don't think I ever will.

Each teacher employed the same simple instructions and, with the aid of a large cardboard model of a Friendship Book page, demonstrated how to place an X in front of the one sentence which best described how an *S* felt about the boy (or girl) whose name appeared at the top of the page. These ratings were then converted to numbers ranging from 5 ("I like him very, very much.") to 1.

High-cohesive (HiCo) and low-cohesive (LoCo) same-sex groups of three or four children were formed on the basis of similar mutual ratings. A single score, indicating the cohesiveness level of a group (its group-cohesiveness index—GCI), was calculated by the formula $\Sigma X/N\ (N - 1)$, where X = a rating, and N = the number of group members who rated each other. This formula takes into account the rating of each group member by every other member and is, thus, an operational measure of our conceptual definition of cohesiveness (Lott & Lott, 1961). Thirty-five HiCo groups were formed, ranging in GCI from 5.0–4.2 (M = 4.59), and 27 LoCo groups were formed, ranging in GCI from 3.3–2.2 (M = 2.91). Of the total of 62 groups, 37 had 3 members and 25 had 4.

In addition to the sociometric ratings made and received, *Ss*' IQ scores were also taken into account in placing them in groups. The *California Short-Form Test of Mental Maturity, Elementary Form, 1957* (CTMM) had been administered several weeks prior to Socio I. A child with a total IQ score of 116 or above was eligible for placement only in a high-IQ (HiI) group; a child with a score of 100 or below was eligible for placement only in a low-IQ (LoI) group. Those with scores between 115 and 101 could be placed in either HiI or LoI groups, but the authors tried to avoid putting an *S* into a LoI group if his IQ score was above 108 and putting an *S* into a HiI group if his score was below 108. It must be kept in mind that *Ss* had to be placed in groups with other children of a compatible IQ level as well as with other children whom they rated highly and were highly rated by (HiCo groups), or whom they rated poorly and were poorly rated by (LoCo groups).

Table 1 describes the independent-variable conditions, HiI-HiCo, HiI-LoCo, LoI-HiCo, and LoI-LoCo groups in terms of number of groups, mean cohesiveness index, total number of *Ss*, and mean individual IQ score. Since some *Ss* (and groups) were present for LTI but absent for the subsequent tasks (RET, REL, and LTII), the independent-variable conditions are also described for the latter. It is clear from Table 1 that *S* losses produced only minor changes in average cohesiveness and IQ-score characteristics.

Learning Task I. The experimenter (*E*), when introduced by the classroom teacher,

TABLE 1

Summary of Experimental Conditions

Condition	No. groups	Mean GCI	Total No. *Ss*	Mean IQ
Learning Task I				
HiI-HiCo	18	4.63	64	120.98
HiI-LoCo	12	2.97	39	120.46
LoI-HiCo	17	4.53	54	99.35
LoI-LoCo	15	2.86	49	96.89
Total	62		206	
Retention, Relearning, and Learning Task II				
HiI-HiCo	14	4.62	47	121.21
HiI-LoCo	10	2.93	28	120.96
LoI-HiCo	15	4.53	47	99.32
LoI-LoCo	10	2.91	33	98.47
Total	49		155	

explained that he wanted to find out how easily children can learn words in a foreign language. The previously formed small groups were then called to the experimental room in a random, predetermined order and given identical task instructions.

LTI consisted of learning Spanish equivalents of English words. *S*s first worked as a group in arranging and assembling the materials: 12 6¼ × 5⅛ inch plastic-covered cards, each containing a four-word Spanish sentence (e.g., "*Este es un vaso.*") and an appropriate india-ink illustration, and 12 plastic-covered 5 × 2 inch cards, each containing a four-word English sentence, equivalent to one of the Spanish sentences (e.g., "This is a glass."). Each English card could be fitted onto a Spanish card by sliding it under a paper clip on top of the latter. *S*s were told that they would each receive three or four cards (depending upon the group size) with a picture and Spanish sentence on each, and were instructed as follows:

> In the center of your table I will place a stack of smaller cards. These will not have pictures on them. Instead, each will have just an English sentence on it....
>
> Your job will be to match up each English card ... with one of the Spanish picture cards.... You can do the matching job best if each one of you, in order, will turn up an English card on the top of the stack, read the sentence out loud, and then see who has the Spanish picture card that goes with it.... [Whoever] has the right Spanish picture card gets the matching English card [and] ... can then put the two cards together....
>
> When you have matched each Spanish picture card with the correct English one, let me know....

The amount of time it took each group to assemble the materials was recorded by an assistant.

Following the group matching procedure, *S*s were told that they would be tested on their knowledge of the English meanings of Spanish words after a period of studying the materials individually. They were instructed to:

> Study your own [cards] ... first. When Miss Briscoe rings this bell, pass your pairs to the ... right. Study [the new cards you've received] ... until the bell rings again. Then pass your cards again to the ... right. We'll keep doing this until you've all had a chance to study all the ... cards twice.... Study quietly....

The interval between bells was either 1½ or 2 minutes, depending upon whether there were four or three *S*s in the group (and thus, upon whether each *S* had three or four cards).

After 12 minutes of study the cards were collected, and a mimeographed answer sheet with 15 spaces on it was distributed to each *S*. Each Spanish test word was presented once, one at a time on a printed card held in front of *S*s for 15 seconds. *S*s were told to write the English meaning or to draw a line in the appropriate space before going on to the next test word.

Retention and Relearning. One week after LTI, *E* reported to the class that they had done well on the previous learning task, and he once more solicited cooperation for additional language learning. The same groups were again called to the experimental room in a predetermined random order where they were given instructions for a retention test on 10 of the 15 Spanish words on which they had been tested in LTI. On the retention test each Spanish test word was exposed on a card to the group for 10 seconds; *E* then turned the card over to expose the English equivalent for approximately 3 seconds, saying the English word aloud as he flipped over the card.

After the RET trial, *S*s were given three relearning trials. The same words were presented by the same technique, but in a new randomly determined order on each trial. *S*s were instructed to stop if they got all 10 words correct on any trial and to sit with the group while the others went on. A new answer booklet was distributed prior to each trial.

Learning Task II. Immediately after the last REL trial, *S*s were introduced to another task, the learning of 10 new Spanish nouns and their English equivalents (e.g., "*silla*"–"chair"). Each *S* was given his own list of the 10 Spanish-English pairs, while the group as a whole received a stack of 10 cards, each with one of the Spanish words on it. *E* suggested the following study procedure:

> Starting with you ... why don't you turn over the first card when I say "Begin." Place it on the table so that everyone can see it and all of you try to find on your list what the Spanish word means in English. Call out the correct English answer as soon as you know it. Then take turns turning up a Spanish card from the center pile. After a few minutes I'll ... give you the same kind of test that I just gave you on the old words.

After *S*s had gone through the stack of cards four times, *E* gave four trials on the new words, using the same procedure as that followed previously for RET and REL.

Sociometric Test II. Approximately 2 weeks after RET, REL, and LTII, a second sociometric test (Socio II) was given in each class by the regular teacher, following the same procedure as that described for Socio I.

RESULTS

Learning Task Scores

The mean scores earned under the four different experimental conditions were calculated for each learning task, separately for each trial. These results are presented in Figure 1. Figures 2 and 3 show mean scores made by high-IQ and low-IQ *S*s, respectively, under HiCo and LoCo conditions, separately for each sex.[3] On the LTI test, the maximum possible score was 15 correct words; the maximum score on all subsequent tests was 10. (In REL and LTII, an *S* who scored 10 on any trial was given the same score on any remaining trials.)

The IQ variable, high versus low, made a significant difference on every learning test, except for the last trial of LTII, with HiI *S*s scoring reliably higher than LoI *S*s in every case. The results of median test analyses using two-tailed probabilities (Siegel, 1956) are as follows: LTI, $\chi^2 = 19.93$, $p < .001$; RET, $\chi^2 = 7.17$, $p < .01$; REL, Trial 1, $\chi^2 = 10.95$, $p < .001$; REL, Trial 2, $\chi^2 = 9.06$, $p < .01$; REL, Trial 3, $\chi^2 = 4.92$, $p < .05$; LTII, Trial 1, $\chi^2 = 10.96$, $p < .001$; LTII, Trial 2, $\chi^2 = 5.63$, $p < .02$; LTII, Trial 3, $\chi^2 = 5.50$, $p < .02$; LTII, Trial 4, $\chi^2 = 3.72$, $p < .10$. Since these findings justify separate treatment of high- and low-IQ *S*s, an analysis of variance for frequency data (Wilson, 1956) was performed separately for HiI and LoI *S*s, for each set of learning scores, to test the effects of cohesiveness and sex. This analysis compares conditions in terms of the number of individuals within each who scored at or above the median and the number who scored below the median. The results of these analyses are given in Table 2.

Examination of Figure 1 will show that in the HiI condition HiCo *S*s did better than LoCo *S*s, as was predicted, in every case except for LTII, Trial 1. The relationship between learning-task scores made by HiI *S*s and cohesiveness is statistically reliable on RET, and on the first two REL trials, and is close to the standard level of significance on the third REL trial, and on the third LTII trial (see Table 2).

The analyses of variance summarized in Table 2 indicate that among HiI *S*s, sex is not reliably associated with the scores (except on LTII, Trial 2, on which the girls did significantly better), and there is little indication of significant cohesiveness-sex interactions. What is not revealed in Table 2, however, but can be clearly seen by inspecting Figure 2, is that the differences in learning between HiCo and LoCo high-IQ *S*s are contributed to primarily by the girls. Separate median tests were performed for girls and boys to evaluate the significance of HiCo-LoCo differences on each learning test. For the high-IQ girls, a significant χ^2 (i.e., $p < .05$) was obtained, indicating a reliable relationship between cohesiveness and learning, on six of the nine measures: LTI ($\chi^2 = 5.97$); RET ($\chi^2 = 5.67$); REL, Trial 1 ($\chi^2 = 3.92$); REL, Trial 2 ($\chi^2 = 5.24$); LTII, Trial 3 ($\chi^2 = 5.67$); and LTII, Trial 4 ($\chi^2 = 5.67$). In contrast, among the high-IQ boys, conditions of high and low cohesive-

---

[3] Preliminary analyses indicated that the number of persons in a group, three or four, had no reliable effect on the learning-task scores so data from both types of groups were combined.

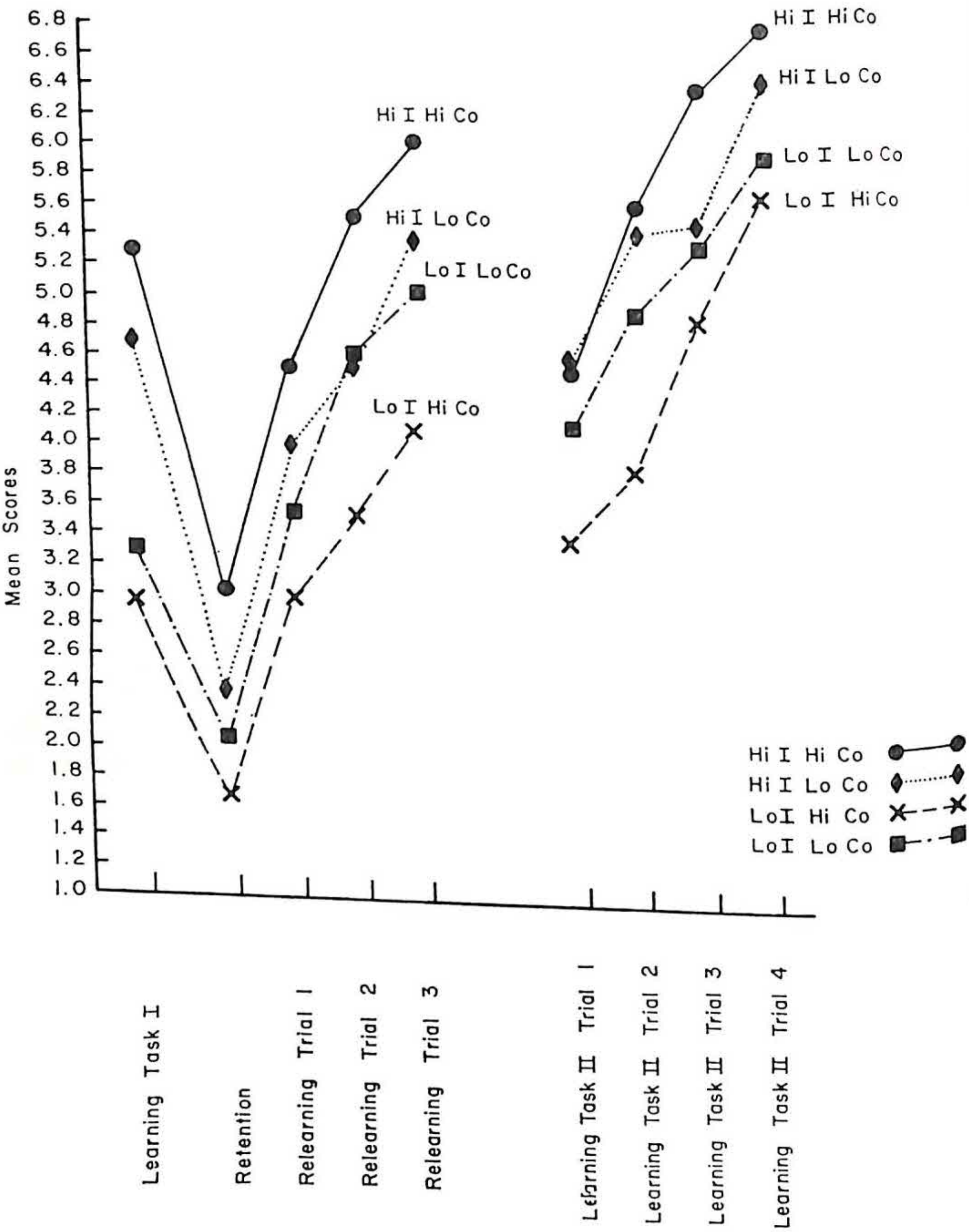


FIG. 1. The experimental conditions (total sample) compared on the learning tasks.

ness made no reliable difference on any learning measure.[4]

Turning now to the low-IQ *S*s, we find a strikingly different pattern of

---

[4] We may look at the same data in another way. On every one of the nine learning measures, HiI-HiCo girls did better, on the average, than HiI-LoCo girls. Using the sign test (Siegel, 1956), we find that the probability that such results are due to chance is .004 (two tailed). This test is not strictly applicable since measures are all

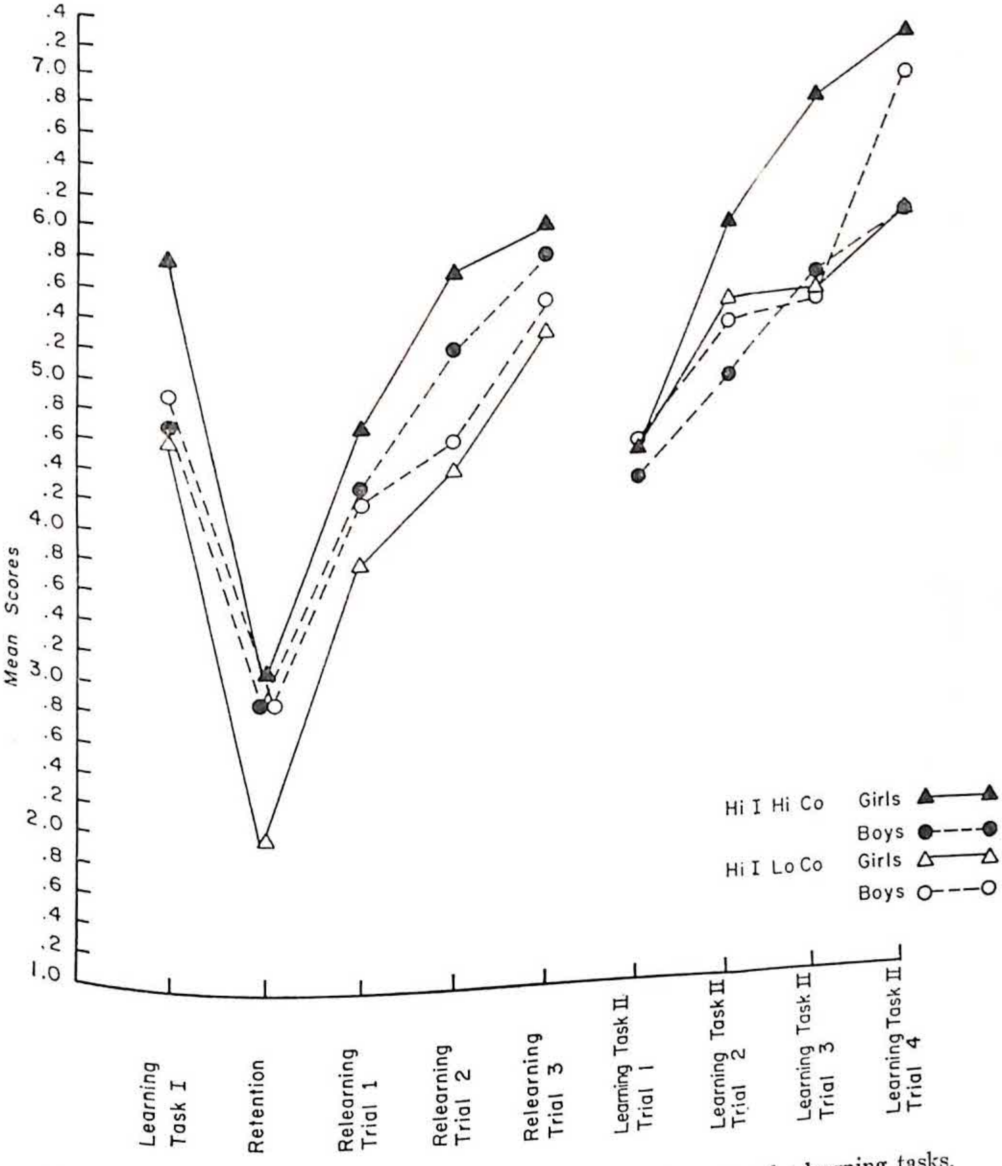


FIG. 2. HiI *Ss* compared, by cohesiveness condition and sex, on the learning tasks.

results. It can be seen in Figure 1 that on every one of the learning-task measures the average score of LoCo *Ss* is greater than the average score of HiCo *Ss*, in direct opposition to the authors' prediction and to the results obtained on HiI *Ss*. The relationship, however, between cohesiveness and the learning measures is not reliable ($p > .05$), as is indicated in Table 2, although significance is approached on RET, REL Trial 2, and LTII Trials 1 and 2.

Separate analyses were performed on the scores made by boys and girls. Among the LoI girls, the average

---

taken on the same *Ss*, but it is suggestive. For the HiI boys, on the other hand, average HiCo scores exceeded LoCo scores on only five out of nine measures; the two-tailed probability that this is a chance phenomenon is 1.00.

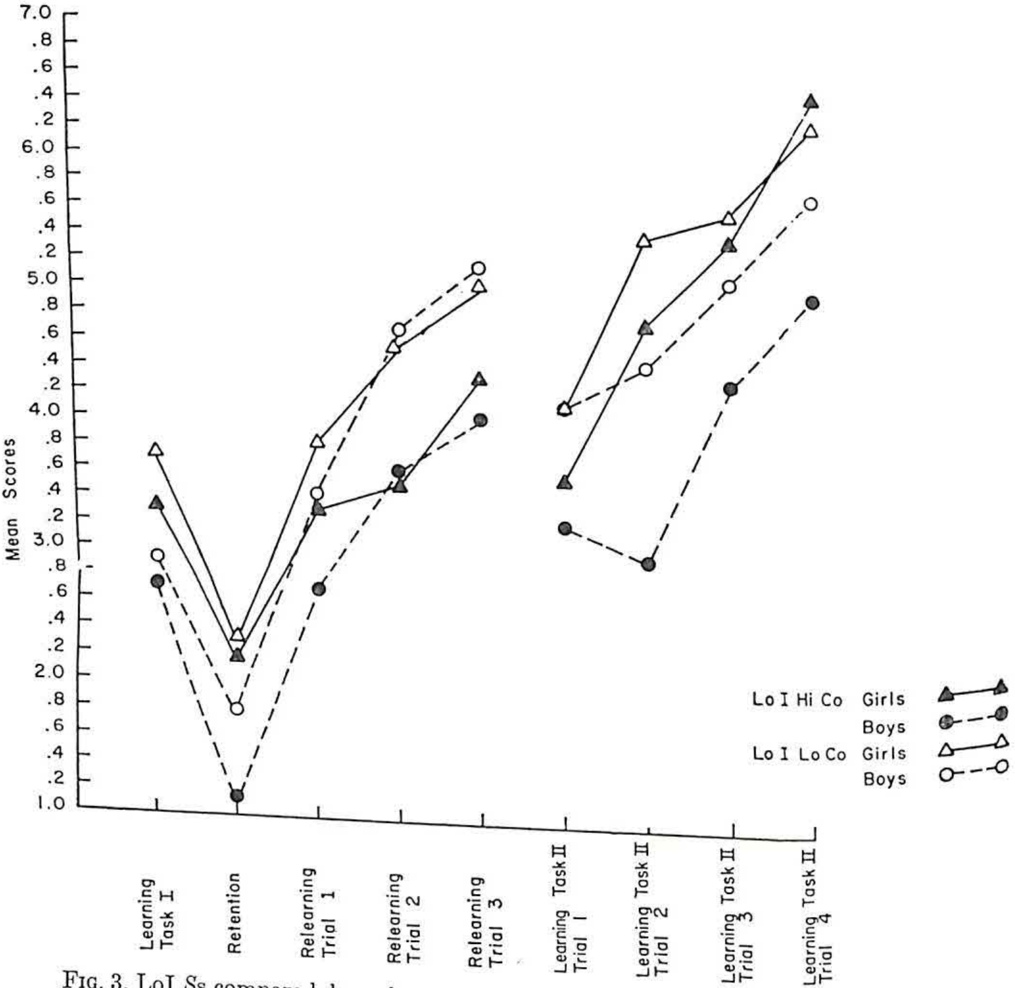


FIG. 3. LoI *S*s compared, by cohesiveness condition and sex, on the learning tasks.

LoCo score exceeds the average HiCo score in every case except for the last trial of LTII, but none of these cohesiveness-score relationships is statistically reliable (median tests). On REL Trial 2, a χ^2 of 2.92 was obtained, which approaches significance ($p < .10 > .05$).[5] For the LoI boys, the average LoCo score is greater than the average HiCo score in every case.[6] Median test analyses of the data on each learning test yielded a χ^2 of 3.36 ($p < .10 > .05$) on RET, and a χ^2 of 6.87 ($p < .01$) on LTII Trial 2. All other chi-squares are nonsignificant.

Time Measure

On LTI a record was kept of the time that it took for each group of children to match the Spanish sentences to their English equivalents. This was a group activity, preliminary to individual study.

The results are shown in Table 3.

---

[5] Using the sign test, it is found that a pattern of eight results out of nine in one direction (LoCo > HiCo) has a two-tailed chance probability of .04.

[6] The probability is .004 that nine out of nine results in one direction is a chance phenomenon (sign test).

TABLE 2

SUMMARY OF ANALYSES OF VARIANCE (FOR FREQUENCY DATA) ON LEARNING-TASK SCORES

Task	HiI *Ss*			LoI *Ss*		
	χ^2 for source of variance[a]			χ^2 for source of variance[a]		
	Cohesiveness	Sex	Interaction	Cohesiveness	Sex	Interaction
LTI	2.47	2.57	3.34*	—	—	—
RET	4.01**	—	—	2.79*	6.13**	—
REL, Trial 1	5.28**	—	—	1.99	—	—
REL, Trial 2	5.07**	—	—	3.45*	—	—
REL, Trial 3	3.33*	—	—	—	—	—
LTII, Trial 1	—	—	—	3.12*	—	—
LTII, Trial 2	2.11	4.25**	1.03	3.09*	2.55	3.83*
LTII, Trial 3	3.54*	—	—	—	—	—
LTII, Trial 4	.43	1.66	2.85*	1.30	2.46	2.25

Note.—For HiI *Ss*, $N = 103$ on LTI, 75 on subsequent tasks; for LoI *Ss*, $N = 103$ on LTI, 80 on subsequent tasks.

[a] $df = 1$ in every case. If there was no possibility of obtaining a significant χ^2 (determined by the magnitude of the total χ^2), further analyses were not performed. This was the case wherever χ^2 figures are not given in the table.

* $p < .10 > .05$.

** $p < .05$.

Differences between means were evaluated with the *t* test (McNemar, 1949). HiCo groups were found to have worked significantly faster than LoCo groups within the total HiI sample, that is, boy and girl groups considered together, and within the HiI sample of girls. The difference between HiCo and LoCo groups of HiI boys was in the same direction, but not statistically reliable. For the LoI groups, the results are quite different; there are no reliable differences between HiCo and LoCo groups within the total sample, or for girls or boys considered separately. In each case, moreover, the LoCo groups were slightly faster, on the average, than the HiCo groups, in reverse of the findings for the HiI groups.

It is of considerable interest that on this time measure, as one would expect, the HiI groups were faster than the LoI groups, on the average, under HiCo conditions, but under LoCo conditions, LoI groups were faster than HiI groups. In neither case is the difference between the IQ conditions statistically reliable ($t = 1.80$, $df = 33$; $t = 1.23$, $df = 25$—for

TABLE 3

TIME TAKEN BY GROUPS TO MATCH SPANISH TO ENGLISH SENTENCES IN LEARNING TASK I

Condition	No. groups	Mean time (in sec.)	*s*	*t*
Total sample				
HiI-HiCo	18	112.78	19.1	2.20*
HiI-LoCo	12	147.33	51.9	
LoI-HiCo	17	135.53	48.5	—
LoI-LoCo	15	127.00	26.3	
Girls				
HiI-HiCo	10	111.00	19.3	2.12*
HiI-LoCo	8	134.12	27.1	
LoI-HiCo	8	130.88	66.3	—
LoI-LoCo	8	117.88	16.6	
Boys				
HiI-HiCo	8	115.00	19.8	1.41
HiI-LoCo	4	173.75	82.3	
LoI-HiCo	9	139.67	28.8	—
LoI-LoCo	7	137.43	32.5	

* $p < .05$ (two-tailed test).

TABLE 4

CHANGES IN AVERAGE RATINGS OF OTHER GROUP MEMBERS FOLLOWING INTERACTION IN THE LEARNING TASKS

Condition	No. *Ss*	Mean ratings on Socio I	Mean ratings on Socio II	M_d	s_d	t
Girls						
HiI-HiCo	30	4.76	4.41	−.36	.08	−4.50***
HiI-LoCo	16	2.97	4.00	1.03	.23	4.48***
LoI-HiCo	21	4.53	4.58	.05	.11	.45
LoI-LoCo	15	2.95	3.74	.79	.29	2.72**
Boys						
HiI-HiCo	18	4.38	4.56	.18	.14	1.29
HiI-LoCo	15	2.84	3.35	.51	.22	2.32**
LoI-HiCo	22	4.53	4.50	−.03	.08	−.38
LoI-LoCo	17	2.95	3.40	.45	.22	2.05*

* $p < .10 > .05$ (two-tailed test).
** $p < .05$ (two-tailed test).
*** $p < .001$ (two-tailed test).

HiCo and LoCo, respectively), but under the HiCo condition the two-tailed probability approaches significance ($p < .10 > .05$).

Changes in Sociometric Ratings

By comparing *Ss*' ratings of their fellow group members on Socio II (subsequent to group interaction) with ratings on Socio I it could be determined whether any changes were related to the experimental conditions.

The data shown in Table 4 are taken from those *Ss* who were present in their groups for RET, REL, and LTII and who were also present when Socio II was given. It can be seen that the children in LoCo groups rated fellow group members significantly more favorably after their group experience than before; this is true of HiI girls and boys, as well as LoI girls and boys. Children who had been in HiCo groups, on the other hand, tended not to change their ratings of one another (which were initially high), except in the case of HiI girls who rated each other on Socio II significantly lower than they had on Socio I. Although the average decrease is small, it is highly reliable.

DISCUSSION

The results on all the learning tests, considered together, indicate that high-IQ children who were in high-cohesive groups did generally better than high-IQ children who were in low-cohesive groups. For low-IQ children, however, cohesiveness, or degree of interpersonal attraction among group members, made no difference, although there was a tendency for such children to do slightly better in low-cohesive than in high-cohesive groups. The findings for high-IQ children thus confirm our hypothesis, while those for low-IQ children do not and suggest an opposite trend. A similar phenomenon was reported by Calvin, Hoffman, and Harden (1957). Trends in three experiments indicated that for *Ss* of high intelligence, group problem solving is better in a permissive than in a traditional group "climate," whereas for *Ss* of average or below-average intelligence, a permissive climate is a handicap.

Post-hoc consideration of our data

has suggested a possible explanation. The authors' predicted that children under high-cohesive conditions would learn better than children under low-cohesive conditions because the former, working in the presence of liked others, would be more highly motivated, or have greater "drive." Since attitudes are presumed to have a drive component, individuals with highly positive attitudes toward other group members should, it was reasoned, have a higher general drive level, other things being equal, than individuals with more neutral or slightly negative attitudes toward their fellow group members. While it is true that the higher the drive, the better the learning for simple tasks where alternative responses are few, it is also known that when a complex task is to be learned, high drive will strengthen incorrect competing responses as well as correct ones, thus impeding the learning process (Farber & Spence, 1953; Taylor, 1956). The authors suggest that this phenomenon has implications for the present experimental conditions. High-IQ children were reliably superior to low-IQ children on all the learning tests, indicating that our tasks were relatively more simple for the former than the latter. Among the former, then, increased drive (assumed to be associated with being in high-cohesive groups) should have improved test performance, while it should have served to depress it among the latter.

Within the high-IQ sample, the cohesiveness conditions made less difference to the boys than to the girls. Why this should be the case is not immediately clear. Perhaps girls have stronger affiliation needs and more intense social ties than boys. Or, the type of verbal learning tasks which the authors employed may be essentially girls' tasks at the developmental period represented by ages 10 and 11. High-IQ girls may be especially at ease in such learning situations, in which case the heightened drive provided by the high-cohesive condition should serve to improve learning for the girls, but not for the boys. Observations of boy and girl groups made by *E* and his assistant during the learning sessions support the view that the former were less at ease and more distractable.

Within the low-IQ sample, the average scores of children in high-cohesive groups tended to be somewhat lower than those of children in low-cohesive groups regardless of sex, although the differences appear to be greater for boys. This is congruent with our assumption that the learning tasks were more difficult (complex) for low-IQ children and for boys.

The results obtained on a time measure in Learning Task I are in line with, and support, the data obtained on the learning tests. HiCo groups were significantly faster than LoCo groups within the high-IQ sample; again, this was especially true of the girls. Within the low-IQ sample, however, the LoCo groups worked faster than the HiCo groups, although not reliably. Thus, for the high-IQ *Ss*, the condition of working with liked others facilitated performance in the sense of increasing task speed, while for the low-IQ *Ss* this same condition seemed to interfere with performance.

The authors' findings on changes in sociometric ratings of fellow group members following interaction in the learning tasks are also interpretable, post hoc, within a learning-theory framework. In all conditions except one the children tended either to increase their ratings of other group members or to rate them the same as they had previously; this is generally expected following neutral or positive interaction. Increased favorability of ratings was statistically reliable for LoCo *S*s, both boys and girls, and both HiI and LoI. Since HiCo *S*s started with ratings of group members which were close to the maximum, the only possibilities following interaction were no change or decreased favorability. The former result was obtained for HiCo boys, both HiI and LoI, and for LoI-HiCo girls. Among HiI-HiCo girls, however, sociometric ratings decreased significantly. In attempting to understand this latter finding, it is important first of all to keep in mind that feedback from the learning tasks was similar for all *S*s. *E* told all *S*s, 1 week after LTI, that they had done well, but no evaluation was given after the other learning tasks. *S*s did know each time they got a word right or wrong on REL and LTII, but they could not compare themselves with other children to determine relative success or failure. The HiI-HiCo girls were superior to all other experimental conditions on every learning measure, but how well they thought they were doing is not known. The fact that their ratings of fellow group members declined may be explainable on the basis of effort, using the concept of reactive inhibition (Hull, 1951). It is reasonable to infer that the HiI-HiCo girls exerted more effort than children in the other experimental conditions both from their superior performance on the learning tests and from the fact that they worked fastest on LTI where a time measure was obtained. It has been suggested (Lott, 1963) that perhaps stimuli (other group members) associated with effort come to evoke avoidance tendencies just as reward-associated stimuli come to signal approach.

In general, findings of the present study can be said to have provided qualified support for the authors' initial hypothesis and to have raised new questions of broader scope for subsequent investigation.

REFERENCES

BINDER, A., MCCONNELL, D., & SJOHOLM, NANCY A. Verbal conditioning as a function of experimenter characteristics. *Journal of Abnormal and Social Psychology*, 1957, **55**, 309–314.

CALVIN, A. D., HOFFMAN, F. K., & HARDEN, E. L. The effect of intelligence and social atmosphere on group problem solving behavior. *Journal of Social Psychology*, 1957, **45**, 61–74.

DOOB, L. W. The behavior of attitudes. *Psychological Review*, 1947, **54**, 135–156.

FARBER, I. E., & SPENCE, K. W. Complex learning and conditioning as a function of anxiety. *Journal of Experimental Psychology*, 1953, **45**, 120–125.

HULL, C. L. *Essentials of behavior*. New Haven: Yale Univer. Press, 1951.

HULL, C. L. *A behavior system*. New Haven: Yale Univer. Press, 1952. Ch. 5.

JAMES, ANNA G., & LOTT, A. J. Reward frequency and the formation of positive attitudes toward group members. *Journal of Social Psychology*, 1964, **62**, 111–115.

LOTT, A. J., & LOTT, BERNICE, E. Group cohesiveness, communication level, and conformity. *Journal of Abnormal and Social Psychology*, 1961, **62**, 408–412.

LOTT, BERNICE E. Group cohesiveness: A learning phenomenon. *Journal of Social Psychology*, 1961, **55**, 275–286.

LOTT, BERNICE E. Secondary reinforcement and effort: Comment on Aronson's "The effect of effort on the attractiveness of rewarded and unrewarded stimuli." *Journal of Abnormal and Social Psychology*, 1963, **67**, 520–522.

LOTT, BERNICE E., & LOTT, A. J. The formation of positive attitudes toward group members. *Journal of Abnormal and Social Psychology*, 1960, **61**, 297–300.

MCNEMAR, Q. *Psychological statistics*. New York: Wiley, 1949.

PERKINS, H. V., JR. The effects of climate and curriculum on group learning. *Journal of Educational Research*, 1950, **44**, 269–286.

REECE, M. M., & WHITMAN, R. N. Expressive movements, warmth, and verbal reinforcement. *Journal of Abnormal and Social Psychology*, 1962, **64**, 234–236.

REHAGE, K. J. A comparison of pupil-teacher planning and teacher-directed procedures in eighth grade social studies classes. *Journal of Educational Research*, 1951, **45**, 111–115.

SAPOLSKY, A. Effect of interpersonal relationships upon verbal conditioning. *Journal of Abnormal and Social Psychology*, 1960, **60**, 241–246.

SHAW, M. E., & SHAW, LILLY MAY. Some effects of sociometric grouping upon learning in a second grade classroom. *Journal of Social Psychology*, 1962, **57**, 453–458.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

TAYLOR, JANET A. Drive theory and manifest anxiety. *Psychological Bulletin*, 1956, **53**, 303–320.

WILSON, K. V. A distribution-free test of analysis of variance hypotheses. *Psychological Bulletin*, 1956, **53**, 96–101.

(Received March 8, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 2, 74–85

EFFECT OF INSTRUCTIONAL SET AND VARIETY OF INSTANCES ON CHILDREN'S LEARNING[1]

HARRIETT AMSTER
Institute of Human Learning, University of California, Berkeley

In experiments with children of approximately 5 and 10 yrs., learning under intentional set was found to be significantly superior to incidental learning in children of both age groups. Moreover, under incidental set older children showed a reliably greater gain due to practice than did the younger ones. Also, for the 10-year-olds, the presence of irrelevant cues inhibited learning and decreased improvement due to practice. No consistent effects were observed for the variety variable, but for both groups there was some trend for the small-variety condition to foster learning when irrelevant features were not included in the stimulus.

The present investigation considers developmental differences in concept formation and memory for concepts. Two contrasting types of responses which could underlie the learning of concepts have been inferred by various investigators (e.g., Inhelder & Piaget, 1958; Vinacke, 1952). Concept formation of the *deductive type* involves use of strategies which might entail deductive reasoning and systematic testing of hypotheses; concept formation of the intuitive or *associative type* entails associative responses to the presented stimuli. The present study bears on the hypothesis that these types of concept learning may differ in prominence depending on the age of the children and on their instructional set.

Set-induced differences in the relative dominance of deductive and associative responding would be expected on the basis of the descriptive research by Inhelder and Piaget (1958) who noted a marked contrast between the intuitive thinking of children aged 4–7 and the operational and logical thinking of young adults. Similarly, others (e.g., Osler & Trautman, 1961) have found experimental evidence in support of similar contrasting processes of thinking occurring differentially as a function of age.

Instructional differences in the relative dominance of deductive and associative responding would be expected on the basis of the present writer's earlier study (Podell, 1958) which suggested that for college students deductive behavior occurs during intentional concept learning while associative responding predominates during incidental concept learning. In the present study two instructional sets were employed, an intentional set in which the subjects (*S*s) were instructed to deliberately form a concept of the similarities among a series of "flags," and an unintentional set in which they were instructed to make aesthetic judgments about the same series of stimuli. There is reason to believe (Podell, 1958) that similar instructions can, at least among college students, induce with differential dominance, deductive and associative responding. It was further hypothesized that 10-year-olds would respond in a manner similar to the college

---

[1] This research was supported by United States Office of Education Projects 1459 and 2243. The manuscript was written under the auspices of the Institute of Human Learning which was supported by a grant from the National Science Foundation.

students, that is, would respond differently under intentional and unintentional instructions, but that 5-year-olds who may not yet have learned deductive strategies would respond associatively under both types of instructions.

The hypothesized differences in type of responding between intentional and incidental instructions may be studied by the observation of their effects on a stimulus variable, the degree of variety in the instances from which the concept is formed. According to the present definition, a large variety would consist of a number of *different* instances of a concept, and the corresponding small variety would consist of an equal number of instances of the concept wherein there are fewer *different* instances; thus, each instance is repeated more frequently in the latter than in the former case. On a logical basis, when *Ss* employ deductive strategies, a large variety should provide an advantage over a small variety because the large variety permits false hypotheses to be rejected in fewer trials than a small variety does (Podell, 1958). However, when *Ss* respond associatively under incidental instructions, there is more question as to the relative superiority of the two degrees of variety. On the one hand, the large variety may permit more efficient summation of correct associations, but on the other hand, in view of the fact that particular instances are repeated more often in a small variety, memory for concept instances should be facilitated by it. Podell (1958) studied these hypotheses in college students and found that under intentional instructions the large variety was, as predicted, more effective in facilitating concept formation than the small variety. However, when unintentional instructions were employed, the small variety was found to be more effective. It would appear that the superiority of the small variety under unintentional instructions stems from the dependence of unintentional concept learning on the ability to recall exemplars of the concept.

The present experiments investigate in 5- and 10-year-old children the learning of concept instances under an intentional and an unintentional set and on the basis of large and small varieties of instances. It was expected that, when given intentional instructions, the older children to whom a large variety was given would excel over those given a small variety, but no corresponding difference would be expected for the younger children. Furthermore, the older children should show better recall of instances on the basis of the intentional set than the unintentional set, but for the younger children, any differences in recall as a function of set should be smaller.

Experiment I

Method

Subjects. The experimental procedures were administered to eight intact classes in the Berkeley (California) Unified School District. Sixty-four kindergarten children and 64 fourth- and fifth-grade children who passed a color vision test were postexperimentally selected as *Ss*. The Science Research Associates (SRA) Primary Mental Abilities Test scales were administered to all *Ss*. The mean mental age (MA) of kindergarten *Ss* was 6 years, 4 months, and the range was 4 years, 10 months to 7 years, nine months. The mean MA of fourth- and fifth-grade *Ss* was 12 years, and the range was 9 years, 2 months to 14 years, 10 months. The chronological age (CA) of kindergarten *Ss* ranged from 4 years, 11 months to 7 years, 2 months, and the CA of fourth- and fifth-grade *Ss* ranged from 9 years, 4 months to 11 years, 9 months.

Experimental Design. For each age group,

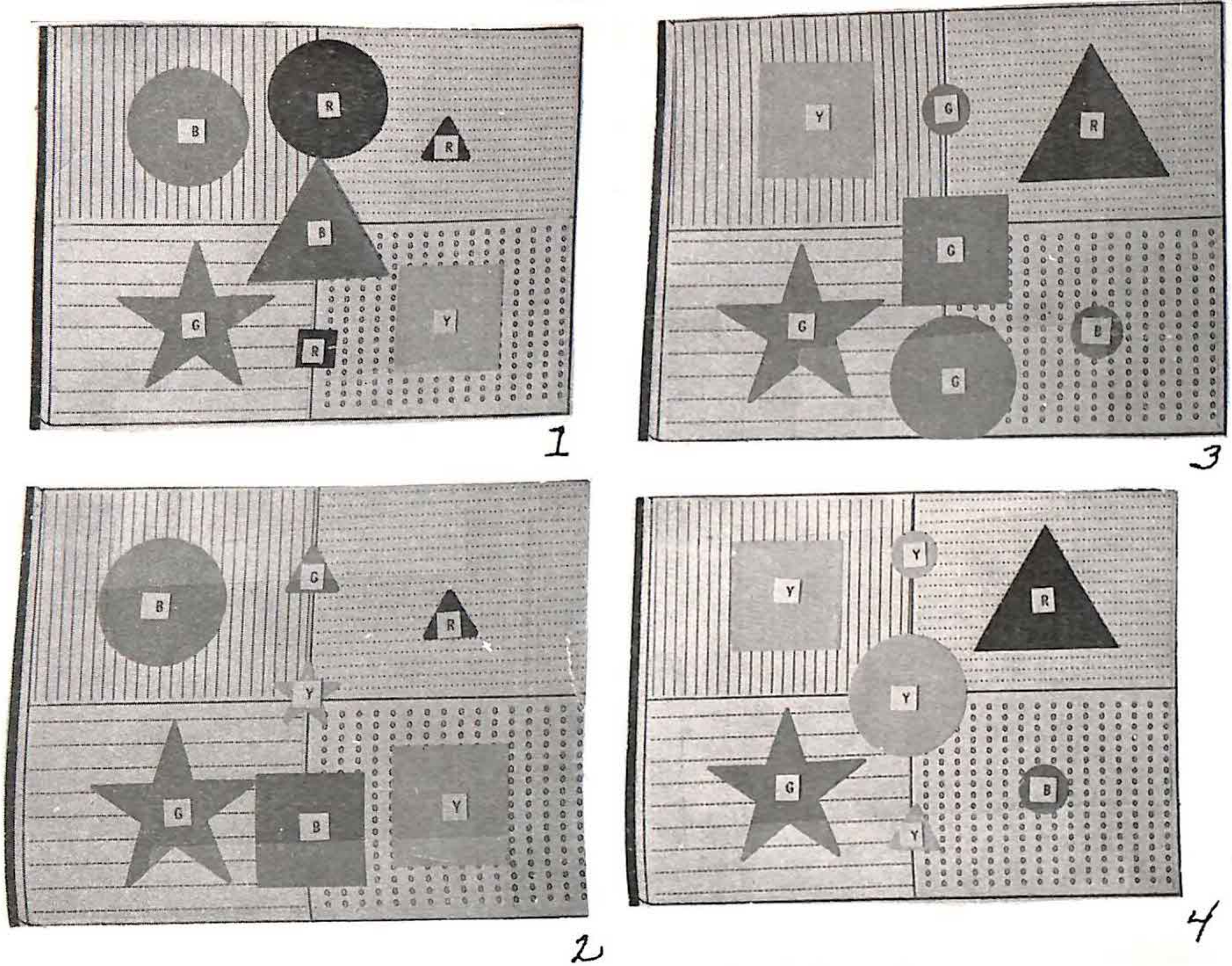


FIG. 1. See next page for continuation of figure.

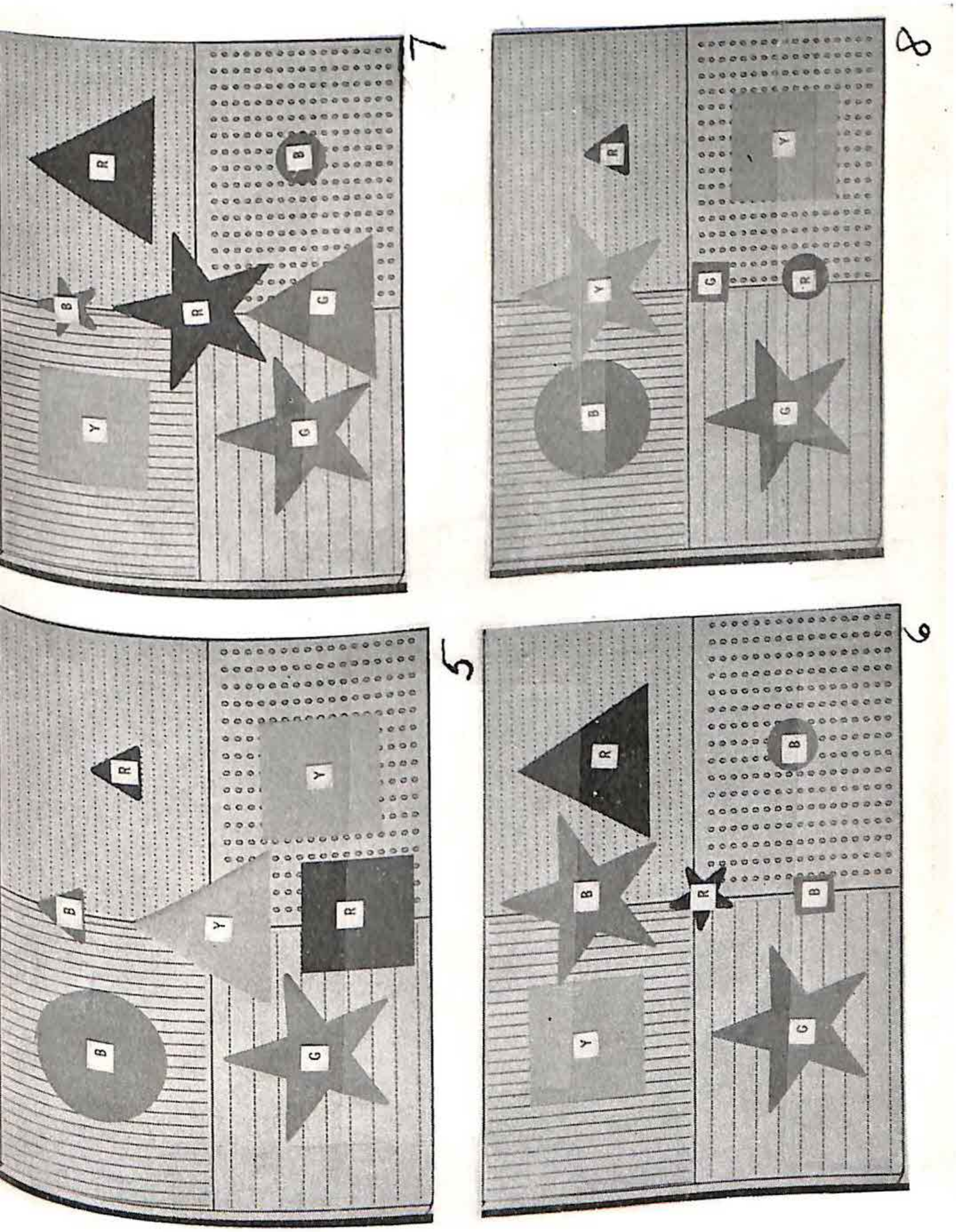


FIG. 1. Sample of the stimuli employed in Experiment II. (A small variety of relevant features and a large variety of irrelevant features.)

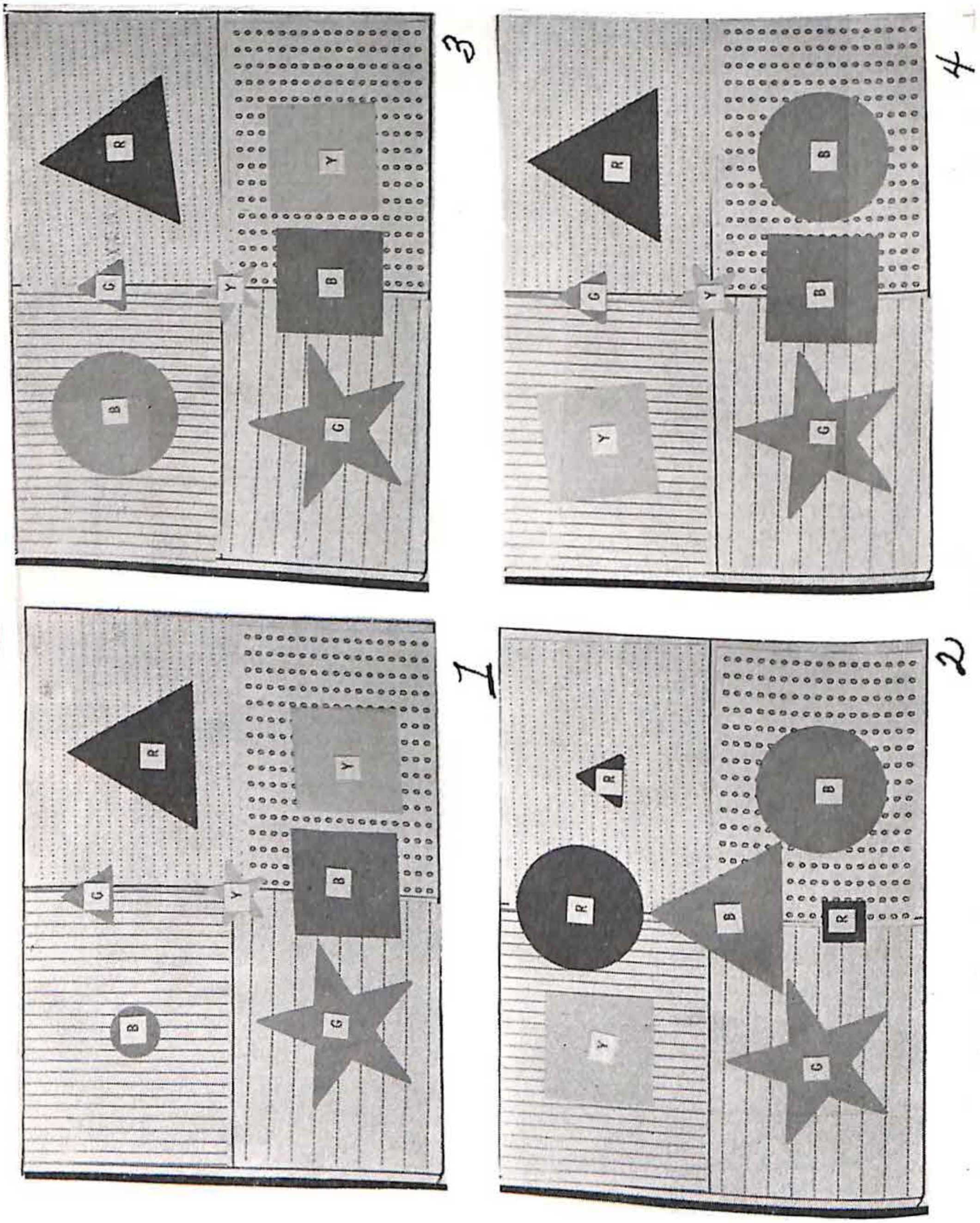


FIG. 2. See next page for continuation of figure.

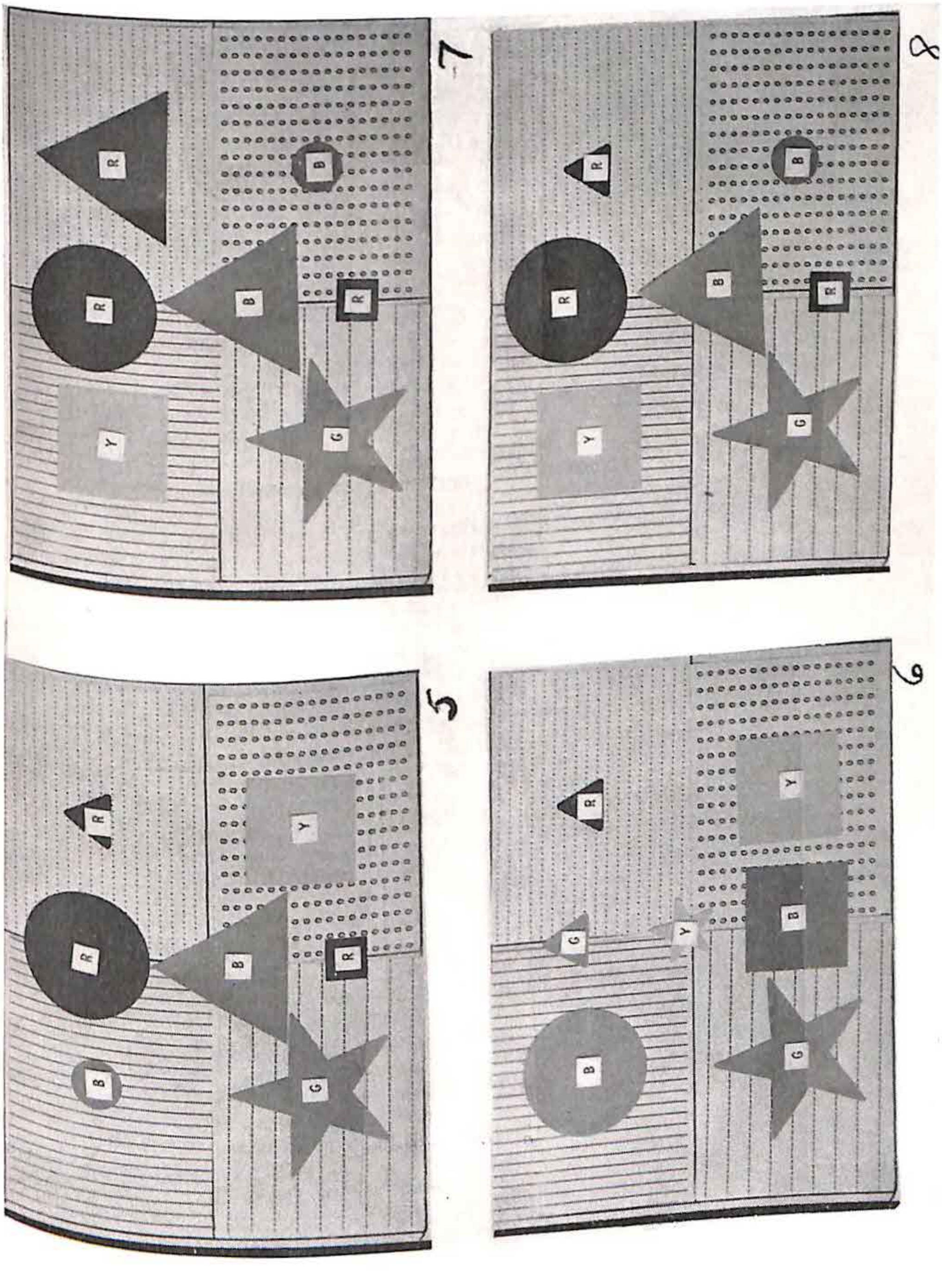


FIG. 2. Sample of the stimuli employed in Experiment II. (A large variety of relevant features and a small variety of irrelevant features.)

there were four major experimental subgroups, corresponding to the 2 × 2 design by which two levels of variety of concept instances and two types of instructional set were manipulated factorially. In addition, within each subgroup *S*s were presented 16 different booklet types. There were 16 different *S*s within each major subgroup at each age level, and each *S* received a different booklet; thus confounding the two sources of error. During the initial testing, many more than the 64 *S*s at each age level were assigned randomly to conditions and tested for color perception. An equal number of *S*s from each subgroup were postexperimentally selected at random from among those who passed this test and to whom the experimental conditions had been administered.

Materials. First to be considered is information relevant to the stimuli presented to an individual *S*. Each *S* received a booklet containing eight pages of flags, a blank page, two response pages, and envelopes containing materials for testing learning. There was only one flag on each of the eight pages. For this experiment, each flag contained four geometric shapes (circle, square, triangle, and star), each in a different color (red, green, yellow, and blue), and each placed in the center of a quadrant. The flags shown in Figures 1 and 2 each contain, in addition, three irrelevant forms in the center column. If these three forms were deleted, Figures 1 and 2 would consist of sample booklets for Experiment I. Ignoring the three irrelevant forms, Figures 1 and 2 each contain the eight flags for one *S*. Factors which were constant across pages within a booklet include the color of each shape (e.g., a square may always have been red) and the pattern on each quadrant (e.g., vertical stripes may always have appeared on the upper right quadrant). In addition, other constant factors varied among shapes. For the shape at Complexity Level 1, both position and size remained constant; for the one at Complexity Level 2, position remained constant but size varied; for the one at Complexity Level 3, position alternated between two quadrants, but size remained constant; for the one at Complexity Level 4, both position and size were varied.

The pages of the booklets were constructed from separate parts—sheets of paper approximately 9 × 11 inches which contained a flag outline, four labels in the size of the quadrants of the flag which had different background patterns, eight shapes (the four shapes each in two sizes) which were die cut from four colors of gum paper. The flags were approximately 7¼ × 5½ inches. The diameter of the large shapes was approximately ¾ inch.

Each *S* was given three envelopes which contained the following materials for testing learning: crayons, each of the four shapes in both sizes cut from white gummed paper, and four labels printed with the appropriate background designs. Each desk was set up before *S*s arrived with a booklet and four envelopes each closed with a rubber band.

Turning next to information relevant to the entire set of stimuli, across the entire set of booklets it is noted that each of the four shapes occurred in two sizes and four colors. The particular assignment of shapes to position, color to shape, and shape to complexity level was made with the aid of a Greco-Latin square. On this basis, 16 different object assignments were made. The background labels were placed on the quadrants in one of two spatial arrangements. Half of the object assignments employed one of the background arrangements and half employed the other. A pair of booklets which differed in the variety of different instances was constructed for each object assignment. One was a small-variety booklet (Figure 1, omitting the three irrelevant forms) in which there was more frequent repetition of the same page than in the large-variety member of the pair (Figure 2, omitting the three irrelevant forms). In all, there were 16 different large-variety booklets and 16 different small-variety booklets which corresponded to the large-variety booklets in respect to the particular assignment of color, form, position, complexity level, and background.

The large-variety booklets consisted of eight *different* instances. For each object assignment there were eight possible ways of making a page, and the pages of the large-variety booklets corresponded to these. Thus, each page was different from every other page, as shown in Figure 2. For the small variety, two different instances were selected from the eight, and each of these was repeated four times. The selection of the two pages was random with the restriction that the two reflected the alternatives required for complexity levels 2, 3, and 4.

Procedure. In general, the procedure involved a study trial in which *S* observed the flags by turning the pages of his own booklet, constructed his own flag, observed the same flags in the same way for a second trial, and, finally, constructed a second flag.

The experimenter (*E*) read one of two sets of instructions and paraphrased them. In the instructions for the intentional set, *Ss* were told that they were to "notice the things which are in all the flags they will see," while in the instructions for the incidental set *Ss* were told to "notice how well you like each one." *Ss* were not permitted to turn pages until instructed to do so as in the administration of the SRA. Ten seconds were given for viewing each page before the instruction to "turn the page" was given. There was no interval between pages other than the 1 second allowed for turning the page. *Ss* were continually reminded to "keep your eyes on your own booklet." The ninth page of the booklet was a blank green sheet. Immediately following the viewing, *Ss* were asked to perform the following tasks in sequence: (*a*) open the small white envelope, (*b*) remove the eight white forms which it contained, (*c*) lay them out on the green page, (*d*) open the small tan envelope and remove the crayons, and (*e*) color the white forms the same color they were in the flags. Next, *S* opened an envelope containing background labels and was told to place four labels on the blank page as if to make a flag. Four of the forms were then to be placed on top of the appropriate label. *S* was to lick the forms so that they would stick to the label, transfer the labels to his desk, turn the page to the blank flag, and stick the labels to which he had attached the forms onto the blank flag. This flag which *S* constructed was collected, and the procedure of viewing and constructing a new flag was repeated.

Screening for Color Blindness or Weakness in Color Identification. After the experiment, colored forms identical to those which had appeared on the flags were distributed to the kindergarten pupils. They were instructed to color an area beneath each form with a crayon of the same hue. The fourth- and fifth-grade pupils were given a similar task and were also shown crayons and asked to write their color names. Pupils making more than one error were not used as *Ss*.

Scoring. The flags constructed by *Ss* were scored by two independent scorers using a standard scoring sheet. A third judge made decisions in the cases on which scorers disagreed. Such instances occurred predominantly among kindergarten *Ss*. Seven categories were scored with respect to each of the possible parts of the flag. These included: the number of forms which were placed anywhere on the flag—maximum correct was four different shapes on one flag—; the number of correct colors placed on the forms—maximum was one form colored each of the four colors—; the correctness of the assignment of color to form—maximum was four forms colored the four colors correctly, regardless of whether they appeared in the correct position—; the correctness of the placement of the forms on the background regardless of quadrant; the correctness of placement of the forms on a quadrant; the correctness of the assignment of color to background regardless of whether the form or quadrant was correct; the correctness of the assignment of color to quadrant regardless of whether form or background was correct. The maximum total score was 28 points.

RESULTS

Analyses of variance and covariance were applied separately to the learning scores for Trials 1 and 2. The covariate, MA, was not found to correlate significantly with the learning scores ($r = .17$ for Trial 1, $r = .22$ for Trial 2). The *F* ratio for regression was 3.49 for the analysis of covariance for Trial 1, and 4.23 for Trial 2 ($p < .10$, $p < .05$, respectively). The analyses of covariance showed only one statistically significant effect—instructional set. As expected, intentional learning was superior to incidental learning for both Trial 1 ($F = 5.27$, $df = 1/95$, $p < .05$) and Trial 2 ($F = 6.80$, $df = 1/95$, $p < .05$). Table 1 presents the mean scores under the two sets as a function of the grade level of *Ss*. Although the effect of Variety did not approach significance, the interaction of Booklet Set and Variety showed a trend ($F = 2.50$, $df = 3/95$, $p < .10$). For booklet types showing the trend at this level of significance, the Small Variety condition was more facilitating than the Large Variety. Although there were no significant interactions between Grade and Set, these data, as shown in Table 1, do suggest that improvement from Trial 1 to Trial 2

TABLE 1

MEAN LEARNING SCORES OF THE TWO AGE GROUPS FOR THE EFFECTS OF VARIETY AND SET BY TRIAL

Experiment I				
Sets and conditions	Trial 1		Trial 2	
	M	*SD*	*M*	*SD*
Intentional set				
Kindergarten	12.31	4.25	14.88	4.67
4th & 5th grade	20.06	6.12	23.50	3.68
Incidental set				
Kindergarten	11.47	5.04	11.90	4.69
4th & 5th grade	16.16	4.95	21.47	5.02
Large variety				
Kindergarten	11.88	4.48	12.88	4.19
4th & 5th grade	17.16	5.11	22.22	4.59
Small variety				
Kindergarten	11.91	4.87	13.91	5.50
4th & 5th grade	19.06	6.71	22.75	5.72

did depend upon Grade and Set in that kindergarten *S*s showed less improvement with practice than did the fourth- and fifth-grade *S*s. Analyses of variance and covariance of the difference in performance between Trials 1 and 2 were also carried out, and the interaction between Set and Grade was significant ($F = 5.47$, $df = 1/95$, $p < .05$).

Analyses of variance and covariance were computed using another score—the number of times color was correctly assigned to forms. The total possible correct for the two trials was eight. As for the other measure, Set was the only variable which showed a statistically significant effect ($F = 7.01$, $df = 1/95$, $p < .01$). The mean scores for the kindergarten *S*s were 2.09 under incidental learning and 3.34 under intentional learning, and for the fourth- and fifth-grade *S*s, 5.03 under incidental learning and 6.22 under intentional learning.

A Friedman two-way analysis of variance was carried out to test the hypothesis that the number of correct assignments of color to forms decreased with complexity level. Instead of performing this analysis for individuals, the four individuals receiving one type of booklet were treated as one individual. Scores were assigned for correct color on form for each complexity level based on the number of persons who assigned the color correctly; these could vary from 0–4 correct. Separate analyses were computed for the two trials and for the kindergarten and fourth and fifth grades. The X_r values ranged from .30–.80 and did not approach statistical significance.

EXPERIMENT II

No evidence for the hypothesized interaction between Instructional Set and Variety of Instances was obtained in Experiment I, and an analysis of the way in which Variety was manipulated suggested that the effect might be observed if a different mode of explicating this variable were employed. The definition of variety already employed was the number of different instances of the concept to be formed in which the instances only contained features relevant to the concept. By contrast, Experiment II included, in addition, a new definition of variety—the number of different irrelevant features within the entire set of exemplars from which an *S* formed the concept. The instances in Experiment I contained only relevant features, the instances in Experiment II contained both relevant and irrelevant features, and the variety of each was manipulated independently by means of a factorial design.

Method

Subjects. The experimental procedures were administered to eight intact fourth- and fifth-grade classes which were roughly comparable in all objective respects to those used in Experiment I. A total of 64 *S*s was included, distributed among 8 groups of 8 *S*s each. The mean MA of the total group was 11 years, 9 months, and they ranged from 9 years, 0 months to 15 years, 7 months. The CAs ranged from 9 years, 3 months to 11 years, 5 months.

Materials. Booklets were prepared which were identical to those used in Experiment I, except that as shown in Figures 1 and 2 three irrelevant forms were added to the center of each page of each booklet. Sixteen booklets were selected from those used in Experiment I of which half were large-variety booklets and half were the corresponding small-variety booklets, and the booklets represented two of the four Greco-Latin squares employed in the earlier study. The 16 booklets were duplicated, making 32 in all. Three irrelevant color forms were added to each page such that there was either a large or a small variety in these forms. One of the two identical booklets was given a small variety of irrelevant forms, and the other was given a large variety.

The irrelevant forms were selected as follows: The pool of possible forms consisted of the 4 shapes in each size and in each color. This yielded 32 possible forms from which the 6 forms used as the relevant forms in a given booklet were excluded. The irrelevant forms were selected from the remaining 26 forms. For the large variety, sets of 3 were selected without replacement for each page. Thus each booklet required 24 of the forms. For the corresponding small variety of irrelevant forms 2 of the sets of 3 were selected at random, and each was repeated on 4 pages of the booklet. Thus there were 32 booklets of 4 types: large or small variety with respect to the relevant forms, and large or small variety with respect to the irrelevant forms.

Procedure. Experiment II was administered in exactly the same way as the first experiment.

TABLE 2

MEAN LEARNING SCORES FOR MAJOR VARIABLES

Experiment II				
Sets and conditions	Trial 1	*SD*	Trial 2	*SD*
Irrelevant variety				
Small	14.69	4.79	16.03	5.13
Large	14.22	4.86	15.06	4.99
Variety				
Small	13.84	4.22	14.66	5.06
Large	15.06	5.30	16.44	4.94
Set				
Intentional	15.31	4.26	16.03	5.41
Incidental	13.59	5.19	15.06	4.68

RESULTS

Analyses of covariance were performed on learning scores with MA as the control variable. Separate analyses for Trials 1 and 2 and the difference scores were computed, but there were no reliable effects. Table 2 presents means of the main effects for the first and second trials. An analysis of variance, mixed design, was performed in which the results of the fourth and fifth grade for Experiment I were combined with those of Experiment II. The variables of this analysis included Presence of Irrelevant Forms, which for Experiment I was Absence of Irrelevant Forms and for Experiment II was Presence of Irrelevant Forms. In addition, Variety, Set, and Trials were employed as in the separate analyses. It reveals that intentional learning was significantly better than incidental learning ($F = 8.33$, $df = 1/120$, $p < .01$), that learning was greater in the absence of irrelevant stimuli than in its presence ($F = 48.00$, $df = 1/120$, $p < .001$), and that improvement occurred from trial to trial ($F = 29.36$, $df = 1/120$, $p < .01$). In addition, the interaction between Presence of Irrelevant Forms and Trials ($F = 10.12$, $df = 1/120$, $p < .01$) indicates that the improvement from trial to trial was greater in the absence of irrelevant forms than in their presence. In the presence of irrelevant forms the mean for Trial 1 was 14.45, and for Trial 2, 15.55. In the absence of irrelevant forms the mean for Trial 1 was 18.28, and for Trial 2, 22.48. The trend of the means for Variety was opposite for the two experiments. However, the interaction between the Presence of Irrelevant Forms and Variety did not reach significance ($F = 2.60$, $df = 1/20$, $p > .05$).

EXPERIMENT III

The effect of the variety of the irrelevant forms was studied in a kindergarten sample without the simultaneous manipulation of the variety of the relevant forms.

Method

Subjects. Kindergarten pupils from two classes of the Berkeley Unified School District served as *Ss*. The 32 pupils preexperimentally selected as *Ss* ranged in age from 4 years, 11 months to 5 years, 11 months. The mean MA was 6 years, 4 months, with a range of 5 years, 5 months to 7 years, 4 months. *Ss* were tested with the SRA intelligence scale and were assigned in advance to the subgroups in the experiment on this basis. The subgroups, numbering 8 *Ss* each, were roughly matched in mean and variance for MA.

Materials. There were 16 different booklets, of which 8 contained a large variety of irrelevant forms, and 8 had a small variety of irrelevant forms. Variety of relevant forms was not manipulated. Eight booklets were selected from among those used in Experiment I. One page which contained four large forms was selected from each of the eight booklets. New booklets were constructed by adding three forms to the one page which was repeated eight times for the eight pages. There were 16 different booklets in all, 8 large-variety and 8 corresponding small-variety booklets. Three irrelevant forms were added to each page as described above for Experiment II. A small variety consisted of two instances from a large variety, each repeated four times.

RESULTS

Analyses of variance were computed for the learning scores of Trials 1 and 2 and for the difference score, but no reliable effects were observed. The main effects are shown in Table 3. Comparison of the scores obtained by kindergarten children in Experiment I with those obtained in Experiment III shows that a decline of about 50% was brought about by the introduction of irrelevant forms despite the fact that there was no variation in the relevant forms for Experiment II. If irrelevant forms were not present, the task for Experiment III would have been easier than for Experiment I; thus, the magnitude of the depression of performance brought about by irrelevant stimuli is, in reality, greater than the observed value.

TABLE 3

SUMMARY OF MEAN LEARNING SCORES

Experiment III				
Sets and conditions	Trial 1	*SD*	Trial 2	*SD*
Variety				
Small	7.81	2.56	8.13	3.20
Large	8.81	1.56	7.63	1.96
Set				
Intentional	8.94	2.02	8.13	2.66
Incidental	7.69	2.15	7.63	2.66

DISCUSSION

As expected, intentional learning was found to be generally more effective than incidental learning. The advantage of intentional learning was hypothesized to be greater for older *Ss* than for younger ones, but no evidence in support of this hypothesis was obtained in Experiment I. However, these data do reveal that kindergarten *Ss* improved with practice under intentional set to a greater degree than under incidental set. In contrast, the older *Ss* showed greater improvement under incidental set than under intentional set. The trend of the means in Experiment II was consistent with these results for fourth- and fifth-grade *Ss*. Quite possibly, under incidental learning, being tested induces a change in the set of the older *Ss*; for example, it alerts them to the need to engage in intentional learning while younger *Ss* may continue to respond on an incidental basis. On the other hand, if 10-year-olds continue to learn incidentally, the effect of practice on incidental learning is greater for them than for younger *Ss*. Also indirect evidence was obtained that the presence of irrelevant forms depresses learning, and the magnitude of this effect was greater for kindergarten *S*s than for older ones.

A clear-cut answer concerning the variety variable was not afforded by the present experiments. While there was a trend for the small variety to facilitate concept learning when the exemplars contained no irrelevant features, this effect was not at all manifest when they were present. A similar trend toward greater facilitation of learning by a small variety has been observed for a somewhat different task (Amster & Marascuilo, 1965). However, this effect has occurred under special circumstances, but has not been consistent over several studies of the acquisition of verbal concepts (Amster, 1965; Podell, 1963) in which the large variety has for the most part been more facilitative.

REFERENCES

AMSTER, HARRIETT. Association as a process in the acquisition of semantic meaning from sentential contexts. Contribution to D. I. Slobin (Chm.), Grammar and verbal behavior in children. Symposium presented at Western Psychological Association, Honolulu, June 1965.

AMSTER, HARRIETT, & MARASCUILO, L. Effect of type of pretraining and variety of instances on children's concept learning. *Journal of Experimental Child Psychology*, 1965, **2**, 192–204.

INHELDER, BARBEL, & PIAGET, J. *The growth of logical thinking from childhood to adolescence.* New York: Basic Books, 1958.

OSLER, SONIA F., & TRAUTMAN, GRACE E. Concept attainment: II. Effect of stimulus complexity upon concept attainment at two levels of intelligence. *Journal of Experimental Psychology*, 1961, **62**, 9–13.

PODELL, HARRIETT A. Two processes of concept formation. *Psychological Monographs*, 1958, **72**(15, Whole No. 468).

PODELL, HARRIETT A. The effect of the variety of instances on the production of verbal concepts. *American Psychologist*, 1963, **18**, 382. (Abstract)

VINACKE, W. E. *The psychology of thinking.* New York: McGraw-Hill, 1952.

(Received March 2, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 2, 86–88

REACTIVE INHIBITION AS RELATED TO PERFORMANCE ON PROGRAMED LEARNING MATERIALS

LOWELL SCHOER
College of Education, University of Iowa

On the basis of performance on a number-canceling task 36 Ss who generated reactive inhibition (RI) quickly and 36 who generated it slowly were selected. All Ss then went through a 1,200-frame programed unit on statistics and measurement. After completing the unit, Ss were administered a test of 36 fact and 36 application-type items. Those Ss who generated RI slowly made significantly more errors on this test than did those who generated RI quickly. Vocabulary-test scores were not significantly related to test performance. None of the interactions among inhibition group, vocabulary, and item type were significant.

Several studies of the role of reactive inhibition (RI) in human learning have resulted from Eysenck's (1952, 1957) incorporation of this construct into his theory of personality. The hypothesis that subjects (*S*s) who generate RI quickly will not learn as efficiently as those who do not has been supported in studies using classical conditioning (Franks, 1957), verbal conditioning (Eysenck, 1959), and reading achievement (Lynn, 1960; Otto, 1963).

The present study was designed to determine whether this hypothesis would be supported when the learning task involved programed materials. The degree to which intelligence and test-item type might influence the effect of susceptibility to RI as related to performance on this type of material was also investigated.

METHOD

Subjects

*S*s were 72 students enrolled in an undergraduate course in educational psychology. These 72 students were chosen from a class of 107, on the basis of performance on a number-canceling task which was used as a measure of RI. Each student was given two 8½ × 11-inch sheets of paper with rows of random numbers on them and was asked to "x" over all the 5's. Total time given was 3 minutes, with *S*s notified each 30 seconds to make a heavy line so that scores per 30 seconds could be determined. At the end of the 3 minutes, all *S*s took a 1-minute break and then canceled for an additional 30 seconds. The total number of numerals inspected during each of the 30-second intervals was then determined for each *S*, and the difference between performance on the first 30 seconds and the 30 seconds just prior to the break was determined as was the difference between the 30 seconds just prior to the break and the 30 seconds after the break. The difference between the score on the first 30 seconds and the 30 seconds just prior to the break was used as the measure of RI. The difference between the scores on the 30 seconds just prior to the break and the 30 seconds just after the break was used to determine whether the task employed to measure RI would show increased scores after a rest interval, and thus provide evidence for the validity of the procedure used to establish the high- and low-inhibition groups.

The 107 *S*s were divided into two groups on the basis of their score on a vocabulary test, with those above the class median put in the high-vocabulary group and those below, in the low-vocabulary group.

In order to control for differences in motivation to cancel numbers and reduce this effect on the results, the placement of *S*s in an inhibition group within each vocabulary group was done in two steps. First, *S*s within each group were ordered on the basis of their canceling scores during the first 30 seconds.

Pairs of *Ss* were then matched as closely as possible on the basis of their score during the first 30 seconds. The member of the pair with the smaller drop from the first 30 seconds to the 30 seconds just prior to the break was put in the low-inhibition group. The member with the larger drop was assigned to the high-inhibition group. Because of problems in matching, only 72 of the 107 students in the class could be used.

This procedure resulted in four groups of *Ss*; a high-vocabulary high-inhibition group (HV-HI), a low-vocabulary low-inhibition group (LV-LI), a high-vocabulary low-inhibition group (HV-LI), and a low-vocabulary high-inhibition group (LV-HI).

Materials and Procedure

The learning material employed in the experiment was a programed unit of approximately 1,200 frames on measurement and statistics. The material was a linear program prepared by the present writer and covered topics such as measures of central tendency and variability, reliability, validity, norms, regression, and prediction. It was presented to *Ss* in textbook form. Each *S* did approximately 80 frames per day for 15 days and was not permitted to study the material outside class. The students' errors on 36 fact and 36 application-type multiple-choice items at the end of the unit were the criterion scores for the experiment.

RESULTS AND DISCUSSION

Two Type I designs (Lindquist, 1953, p. 267) were used to determine the validity of the procedure used to establish the high- and low-inhibition groups. Table 1 shows the mean scores of the high- and low-inhibition *Ss* on the first 30 seconds and the 30 seconds just prior to the break. The main effect for interval was significant at well beyond the .01 level ($F = 103.59$, $df = 1/70$). The interaction between inhibition group and interval was also significant at well beyond the .01 level ($F = 22.09$, $df = 1/70$). This suggests that there was a significant drop in performance across the 3-minute period, and that the drop tended to be different for the high- and low-inhibition *Ss*.

TABLE 1

MEAN NUMERALS INSPECTED BY HIGH- AND LOW-INHIBITION *Ss* DURING THE FIRST 30 SECONDS AND 30 SECONDS PRIOR TO THE BREAK

Ss	First 30 sec.	Last 30 sec.
High-inhibition	129.6	99.0
Low-inhibition	128.8	117.6

TABLE 2

MEAN NUMERALS INSPECTED BY HIGH- AND LOW-INHIBITION *Ss* DURING THE INTERVAL IMMEDIATELY PRECEDING AND THAT IMMEDIATELY FOLLOWING THE REST PERIOD

Ss	Preceding	Following
High-inhibition	99.0	110.8
Low-inhibition	117.6	128.2

TABLE 3

MEAN ERROR SCORES OF HIGH- AND LOW-INHIBITION *Ss* ON 36 FACT AND 36 APPLICATION-TYPE ITEMS BY VOCABULARY LEVEL

Ss	High vocabulary		Low Vocabulary	
	Fact	Application	Fact	Application
High-inhibition	3.94	8.61	4.44	10.22
Low-inhibition	4.39	9.33	6.06	11.61

Note.—$N = 18$ per cell.

RI dissipates during a rest period, so a test of the validity of the procedure used to set up the two groups of *Ss* was a comparison of performance during the interval immediately before and the interval immediately following the rest period. The mean scores of the two groups of *Ss* during these two intervals are shown in Table 2. The difference between the intervals is significant at well beyond the

.01 level ($F = 43.85$, $df = 1/70$). It should perhaps be noted that reminiscence did not differ for the two groups, being 11.8 for the high group and 10.6 for the low group. This could be interpreted to suggest that the rate of dissipation is independent of amount of RI accumulated, but no direct test of this hypothesis was attempted. This finding is, however, contrary to what would be expected from RI theory.

A Type VI design (Lindquist, 1953, p. 292) was used to analyze the error scores on the test taken by *S*s at the end of the unit. Vocabulary was treated as a between factor. Item type and inhibition group (because of the matching involved) were treated as within factors. The mean error scores for the various groups are given in Table 3. The F for vocabulary was not significant, nor were the Fs for any of the interactions involving vocabulary. The F for item type was significant at well beyond the .01 level ($F = 183.81$, $df = 1/34$), but none of the interactions involving item type were significant. The F for inhibition group was significant at greater than the .025 level ($F = 6.25$, $df = 1/34$). Inspection of Table 3 indicates that it was the high-inhibition group, not the low-inhibition group, which made the fewest errors, and this was true regardless of item type or vocabulary level involved. This effect is, of course, the exact reverse of that found in earlier studies using other types of learning tasks. This finding, plus the nonsignificant effect of vocabulary-test score, suggests that when programing materials are involved, the influence of certain variables, such as intelligence, that have a significant impact on performance when other methods of instruction are employed, may not have a significant impact when programed materials are used. It further suggests that certain variables such as RI which have not been extensively investigated in relation to classroom performance may have a significant impact when programed materials are employed.

No very ready explanation for the direction of the effect of RI in the present experiment suggests itself. It had been anticipated that because programed learning presumably involves operant conditioning, students who generate RI quickly would do less well than those who generate it slowly. The results of the present study when compared with earlier studies might suggest that the generation of RI as measured in the present experiment is related to a learner characteristic that has very different effects as type of learning material shifts. In this way it may operate somewhat like anxiety which may well facilitate learning of a simple task, but has a negative effect on learning of a difficult task.

REFERENCES

EYSENCK, H. J. *The scientific study of personality.* London: Routledge & Kegan Paul, 1952.

EYSENCK, H. J. *Dynamics of anxiety and hysteria.* London: Routledge & Kegan Paul, 1957.

EYSENCK, H. J. Personality and problem solving. *Psychological Reports,* 1959, **5,** 592.

FRANKS, C. M. Personality factors and the rate of conditioning. *British Journal of Psychology,* 1957, **48,** 119-126.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

LYNN, R. Individual differences in introversion-extraversion, reactive inhibition, and reading attainment. *Journal of Educational Psychology,* 1960, **51,** 318-321.

OTTO, W., & FREDERICKS, R. C. Relationship of reactive inhibition to reading skill attainment. *Journal of Educational Psychology,* 1963, **54,** 227-230.

(Received November 16, 1964)

Journal of Educational Psychology
1966, Vol. 57, No. 2, 89–98

CHANGES IN ATTITUDES AND VALUES ASSOCIATED WITH COLLEGE ATTENDANCE[1]

IRVIN J. LEHMANN
Michigan State University

BIRENDRA K. SINHA
Sandusky Valley Guidance Center, Ohio

AND RODNEY T. HARTNETT
University of South Florida

To investigate the relationship between length of college attendance and changes in stereotypic beliefs, dogmatism, and value orientation, a battery of instruments was administered to 1,747 Ss as freshmen and then 4 years later whether or not they were still in attendance. The principal findings were that regardless of sex and length of college attendance (a) all groups became less stereotypic in their beliefs, less dogmatic, and with few exceptions, more "outer directed" in their value orientation, and (b) there was, for all practical purposes, no significant relationship between general academic aptitude and either the degree or direction of personality changes. Female Ss, regardless of time spent at college, underwent a more marked change in their attitudes and values between 1958 and 1962 than did their male counterparts. Hypotheses were advanced to explain some of the findings.

In a previous paper (Lehmann, 1963), it was demonstrated that there are marked changes in the critical thinking ability, attitudes of stereotypy and dogmatism, and traditional value orientation of college students between their freshman and senior years. It was also seen that, although some change occurred during each of the 4 years, the most dramatic changes took place during the freshman and sophomore years. These findings suggested that previous research findings that were based upon a time interval of 4 years, where only a pre- and posttest were administered, could be erroneously interpreted.

The study of attitudes and values, especially those relating to change, is beset by a variety of problems not encountered in cross-sectional research. Specifically, some of these problems relate to the reliability of difference scores, measurement errors and regression phenomena, tests of significance, and multiple comparisons, to name a few.

Although it is sometimes asserted that attitudes and values are instilled early in life and are most easily modifiable in infancy and adolescence, curriculum planners in our colleges and universities assume that attitudes and values are still modifiable at ages 17–22, or older. Furthermore, college faculties believe their institutions provide experiences which are unique and are not found outside the academic environment.

Contradictory evidence has been presented regarding the impact of college on student attitudes and values (Jacob, 1957; Lehmann & Dressel, 1963; Newcomb, 1943; Webster, 1958). At the same time, data have indicated that from the freshman to senior years changes in attitudes and values will occur (Jacob, 1957; Lehmann & Dressel, 1963; Newcomb, 1943; Plant, 1962; Webster, Freedman, & Heist, 1962), but that the degree and extent to which attitudes and values are modifiable depend upon the nature of the experience (Smith, 1955), the type and nature of contact (McGuigan, 1958), the personality makeup of the individual (Helson, Blake, Mouton, & Olmstead, 1956), the group's approval of new attitudes (Rosenberg, 1956), and the subject's perception of the outcome (Carlson, 1956). Although the importance of the college climate in changing attitudes and values of college students is apparent in the works of Eddy (1957, 1959), Brown and Bystrym (1962), Lazure (1959), and Newcomb (1962), it has been found to be nearly impossible to single out any one factor as being responsible for college students' attitudes and value changes (Lehmann & Payne, 1963).

---

[1] The research reported in this article was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare.

One of the major faults of many so-called longitudinal studies has been a lack of any periodic study of the original population between the initial and final stages. As Corey (1936) pointed out nearly 3 decades ago, the only valid manner in which to study changes in college students' attitudes and values is by the repeated testing of the *same* students as well as repeated testings with students not in college. Barton (1959) recognizes the difficulty of having an appropriate control group when he comments that the comparative panel study, although involving problems, is the best design unless we are willing to randomly give or withhold a college education to some groups of young people.

Although this study does not claim to have considered or circumvented all of the problems involved in longitudinal research (e.g., absence of a true control group or problems of selective mortality), an attempt has been made to follow up an original population who entered college in 1958, some of whom did not attend for 4 years.

The purpose of this study was to investigate the relationship between amount of college education and changes in attitudes of stereotypy and dogmatism, in traditional-value orientation, and in certain selected attitudes and views related to higher education. In addition, we were interested in studying (*a*) the relationship between changes in selected personality characteristics and intelligence, and (*b*) whether one sex changed more markedly in their attitudes and values than the other. In substance, we were interested in learning what effect higher education has upon the personality development of students and, on the other hand, whether a noncollege atmosphere will produce changes in the same direction and to the same degree as does a college atmosphere.

We do not attempt to identify those factors which might have an impact upon student personality development. Nor are we concerned here with learning whether or not college facilitates the development of certain behavioral traits.

INSTRUMENTS

The Inventory of Beliefs (American Council of Education, 1950) was used to measure stereotypic beliefs. A high scorer is thought to be mature, flexible, adaptive, and democratic in his relationships with others; a low scorer, immature, rigid, compulsive, and authoritarian in his relationships with others.

The Differential Values Inventory (Prince, 1957) was used to measure the subjects' value orientation. This instrument places subjects' values along a traditional-emergent continuum. Using Riesman's classification, one might call the traditional values "inner-directed" and the emergent values "outer-directed" or "other-directed."

The Dogmatism Scale (Rokeach, 1960) measured subjects' receptivity to new ideas, their degree of open-mindedness, and their authoritarian attitude. High scorers are thought to be dogmatic, closed-minded, and unreceptive to new ideas.

An Experience Inventory (Lehmann & Dressel, 1963) was designed to measure social and political attitudes as well as attitudes towards the objectives of college (vocational versus liberal arts), factors important for gaining prestige with students and faculty, and general impressions of college life.

SAMPLE

The experimental group was originally to have consisted of those students maintaining continuous enrollment in college over a 4-year period. It was soon obvious, however, that many students were withdrawing (more specificially, failing to reenroll) for 1 or 2 terms, after which they would return to college. It seemed unrealistic to eliminate these students from the study since their exposure to college—in this case, the "treatment"— was probably no less pervasive than that of students enrolling for each of the 12 terms continuously. In addition, we were guided by the fact that these "withdrawals" were, for the most part, leaving after their sophomore year, and previous research (Lehmann & Dressel, 1962) demonstrated that the most marked changes in attitudes and values occurred during the freshman year. Consequently, the experimental group —or "4-year group" as we shall call it—consisted of those students who entered Michigan State University in Fall, 1958, were in attendance at that institution for at least 9 of the next 11 terms, and were registered as students during spring term, 1962.

The control group consisted of those students who enrolled as freshmen in Fall, 1958, withdrew from MSU, and were not enrolled as full-time students at MSU in Spring, 1962. Students who transferred to another college were considered as being in the control group unless they completed the equivalent of 10 terms of college.[2] The control group was divided, for purposes of analysis, into three subgroups. Group 1C consisted of those subjects who had less than 1 year of college attendance; Group 2C consisted of subjects who had more than 1 but less than 2 years of college; and Group 3C consisted of subjects who had more than 2 but less than 3 years of college.

METHOD

The instruments previously described, with the exception of the Experience Inventory, were administered during orientation week to all freshmen (1,436 males and 1,310 females) entering Michigan State University in Fall, 1958. The subjects in the experimental group were then retested near the end of their senior year, while those in the control group were retested at about the same time by means of a mailed questionnaire. The analysis of data is restricted to those students for whom we have usable pre- and postdata.

TABLE 1

POSTTEST SAMPLE CLASSIFIED BY SEX AND AMOUNT OF COLLEGE EDUCATION

Group	Sample size		No. returning usable test materials		% Return	
	M	F	M	F	M	F
Experimental	966	705	650	484	67	69
1C	217	243	110	147	51	60
2C	112	212	61	104	54	49
3C	141	150	90	101	64	67
Total	1436	1310	941	836	66	64

Because of initial sex differences, the data for the males and females were analyzed separately. To test the hypothesis that there is no difference in the degree and direction of change in attitudes and values among individuals who attended college for varying lengths of time, analysis of covariance was employed where initial (1958) differences among the various educational groups were controlled. After the appropriate *F* was computed, the Scheffé (1959) method of post-hoc comparisons was employed in an attempt to ascertain where the differences actually existed.

To test the hypothesis that there is no relationship between changes in attitudes and values and general academic aptitude, the Pearson product-moment coefficient was computed. Using the 1958 and 1962 test scores (initial and final, respectively), regression equations were computed separately for the males and females in the experimental group and for each of the three control groups. These regression equations were then used to predict the final (1962) scores for stereotypic beliefs, dogmatism, and traditional-value orientation. To compute the amount of change between 1958 and 1962, the difference between the actual pretest (1958) and predicted postest (1962) score was computed for each subject for each of the three variables. These change scores were then correlated with the College Qualification Test total score obtained in 1958 in an attempt to learn whether general academic aptitude is related to certain personality changes.

---

[2] Those *Ss* completing 10 terms were considered as part of the experimental group.

To test the hypothesis that there is no sex difference in magnitude of change in attitudes and values between 1958 and 1962, analysis of covariance was used with the initial (1958) test score being controlled.

RESULTS AND DISCUSSION

The data relative to changes in attitudes of stereotypy and dogmatism, and traditional-value orientation among the four educational groups are presented in Tables 2–4. Table 5 presents the data pertaining to the relationship between general academic aptitude and change. The data concerning sex differences in degree and direction of change are presented in Table 6.

Reference to Table 2 indicates that for both males and females there was a marked change in attitudes and values between 1958 and 1962, regardless of the amount of time spent at college. The degree and direction of changes might be summarized as follows:

1. There is a significant difference among the four groups of males in the amount of change in traditional-value orientation. Whereas those males who had 3 years or less of college became more traditional-value oriented between 1958 and 1962, their counterparts who completed 4 years of college became more emergent in their value orientations between their freshman and senior years.

It should further be noted that the males in Groups 1C and E underwent marked value changes between 1958 and 1962. The difference between the groups stemmed not so much from the magnitude of change within each group, but from the fact that they moved in opposite directions in their value orientations (a mean change of +1.94 and −2.40 points, respectively).

TABLE 2

MEANS, STANDARD DEVIATIONS, *F* RATIOS FOR CHANGES IN ATTITUDES AND VALUES BETWEEN 1958 AND 1962

Measure	Level of education group means								*F*
	1C		2C		3C		E		
	1958	1962	1958	1962	1958	1962	1958	1962	
Males									
Inventory of Beliefs	57.03 (13.63)	65.82 (12.72)	61.03 (14.65)	72.31 (13.22)	63.09 (13.37)	72.13 (14.67)	64.64 (14.41)	72.93 (15.02)	1.33
Differential Values Inventory	34.70 (7.43)	36.64 (8.98)	35.74 (6.72)	35.82 (8.08)	33.78 (6.84)	34.90 (7.39)	34.81 (7.04)	32.41 (7.63)	19.80*
Dogmatism Scale	172.49 (24.84)	155.79 (27.99)	172.26 (27.03)	151.92 (24.34)	167.14 (23.88)	152.42 (23.98)	166.63 (25.54)	154.31 (22.61)	1.29
Females									
Inventory of Beliefs	62.72 (11.73)	68.56 (13.05)	64.15 (14.28)	72.34 (13.87)	64.23 (12.22)	73.93 (13.47)	65.95 (12.92)	78.64 (14.86)	16.51*
Differential Values Inventory	34.38 (7.23)	35.67 (5.92)	32.39 (6.91)	32.80 (7.42)	34.51 (7.02)	32.68 (7.42)	34.24 (6.87)	29.63 (6.79)	30.39*
Dogmatism Scale	168.68 (25.90)	151.49 (23.55)	164.09 (28.85)	148.30 (23.21)	163.94 (23.88)	149.23 (24.08)	163.11 (24.70)	146.99 (23.97)	0.57

Note.—Standard deviations appear in parentheses. *F* is analysis of covariance *F* with 1958 scores used as covariate.
* $p < .05$.

TABLE 3

SCHEFFÉ POST-HOC COMPARISONS TEST OF DIFFERENTIAL VALUES SCORES FOR VARIOUS GROUPS OF MALES AND FEMALES

Groups compared	M	S_D	t^a
Males			
1C and 2C versus 3C and E	37.05 32.92	0.74	5.58*
1C versus 2C	37.97 36.18	0.46	3.89*
1C versus E	37.97 32.38	0.88	6.35*
2C versus E	36.18 32.38	0.86	4.42*
Females			
1C and 2C versus 3C and E	33.91 30.41	0.20	17.50*
1C versus 2C	35.26 32.54	0.55	4.95*
1C versus E	35.26 29.63	0.62	9.08*
2C versus E	32.54 29.63	0.64	4.55*

[a] $t = \frac{\bar{X}_A - \bar{X}_B}{S_D}$.

* $p < .05$. This value is not the conventional table value with $n1$ and $n2$ df, but is computed a priori by a method found in Scheffé (1959).

In other words, those males in Group 1C moved, as a group, towards the traditional-value orientation whereas those males in Group E moved towards the emergent-value end of the continuum.

2. For the females, there is a significant relationship between amount of college education and change in value orientation. There was some difference between the sexes in respect to the direction of change in values. While those males who had either 1, 2, or 3 years of college became more traditional-value oriented between 1958 and 1962, this was not the case for their female counterparts—only those females who withdrew during their freshman or sophomore year became more traditional-value oriented. While those males who withdrew from college during their junior year became more inner-directed, their female counterparts became more outer- or other-directed.

It is conceivable, of course, that the changes were different for the males and females because of the differences in their experiences. This would be especially true if many of the "withdrawal" females married, resulting in their universe of experiences becoming more limited than those of their male counterparts.

3. For both males and females, there is no significant relationship between amount of college education and changes in dogmatism, receptivity to new ideas, or an attitude of open-mindedness. Although all four groups of males and females became less dogmatic between 1958 and 1962, there

TABLE 4

SCHEFFÉ POST-HOC COMPARISONS TEST OF INVENTORY OF BELIEFS SCORES FOR VARIOUS GROUPS OF FEMALES

Groups compared	M	S_D	t^a
1C and 2C versus 3C and E	70.45 77.39	1.20	5.78*
2C versus E	72.38 78.68	1.33	4.66*
1C versus 2C	68.56 72.38	0.96	4.00*
1C versus E	68.56 78.58	1.32	7.59*

[a] $t = \frac{\bar{X}_A - \bar{X}_B}{S_D}$.

* $p < .05$. This value is not the conventional table value with $n1$ and $n2$ df, but is computed a priori by a method found in Scheffé (1959).

TABLE 5

CORRELATIONS BETWEEN CQT TOTAL SCORE AND CHANGE SCORES FOR MALES AND FEMALES ON THE INVENTORY OF BELIEFS, DIFFERENTIAL VALUES INVENTORY, AND DOGMATISM SCALE

Amount of college attendance	Correlations between CQT total score and			
	N	IB change score	DVI change score	Dogmatism change score
Less than 1 yr. (1C)	61 (102)	.09 (.07)	−.01 (−.11)	−.15 (.01)
4–6 terms (2C)	64 (101)	−.33** (.00)	−.09 (.07)	.29* (.12)
7–10 terms (3C)	135 (150)	.06 (.11)	−.13 (−.02)	.06 (−.04)
Seniors (E)	562 (436)	−.01 (.00)	.02 (−.01)	.02 (.13)

Note.—First set of figures for males; those below in parentheses for females.

* $p < .05$.

** $p < .01$.

was no one group that changed more markedly than another.

4. For the females, but not for the males, there is a significant relationship between amount of college attendance and decrease in stereotypic beliefs. In other words, although both males and females in the various groups became less authoritarian and more adaptive and flexible between 1958 and 1962, no one male group changed more markedly than another, although the converse was true for the females. Specifically, the longer a female attended college, the greater the tendency was for her to become less stereotypic in her beliefs. In fact, the magnitude of change occurred in a somewhat linear fashion.

TABLE 6

SUMMARY OF ANALYSIS OF COVARIANCE FOR POSTTEST SEX DIFFERENCES IN STEREOTYPIC BELIEFS, TRADITIONAL VALUES, AND DOGMATISM USING PRETEST SCORE AS COVARIATE

Measure	F ratios for group			
	1C	2C	3C	E
Inventory of Beliefs	0.10	1.10	0.30	37.71***
Differential Values Inventory	5.24*	1.68	11.65***	35.05***
Dogmatism Scale	0.45	0.00	0.13	17.42***

* $p < .05$.

*** $p < .001$.

It is of further interest to note that a temporary plateau was reached between the female sophomore- and junior-year withdrawals.

It should be readily evident that it is very difficult, if not impossible, to isolate a unique factor that would explain attitude and value changes. Because of the impact and continual interaction of a variety of forces—academic experiences, nonacademic experiences, the personality makeup of the individual, and so forth—the researchers must conclude that there are changes in certain attitudes and values associated with college attendance. But they find no compelling evidence which leads them to isolate a particular cause.

5. In those instances where the obtained F was not significant, thus signifying that there was no significant difference among the four college-attendance groups in a particular trait after 4 years (e.g., for the males and females there was no significant change in dogmatism associated with amount of college education), the direction of change was the same for all groups. Combining this with the previous findings, it is seen that regardless of the amount of time an individual attends college, the change manifested over 4 years is toward a lessening of stereotypic beliefs and an increase in receptivity to new ideas. Needless to say, the change was more marked for some groups than for others. In the area of value change, however, there was no

clear-cut pattern evident in the direction of change for the various groups. Further, the males differed from the females in many respects. The only similarities in the direction of value changes for the sexes were for those who attended college for less than 1 year (became more traditional-value oriented between 1958 and 1962), and those students who attended college for 4 years (became more emergent-value oriented).

These findings appear to be contradictory to some of those presented by Jacob and in agreement with those of Plant. Whereas Jacob (1957) concluded that college graduates appear to manifest a system of behavior that is characteristic of college graduates, we have seen that some of the withdrawal subjects manifest the same type of behavior as do the college graduates in some instances. Although Jacob concluded that neither courses, instructors, nor method of teaching have any marked impact upon student values—and it might appear that he was correct when one looks at *some* of the findings presented here—the results of this and other studies (Hall, 1950; Lehmann & Dressel, 1962) do not appear to support this view completely. All in all, it would appear that college acts as a catalyst to speed up changes that would ordinarily occur as the individual matures (Plant, 1962). Even if this were the only impact that college has upon student attitudes and values, its importance should not be minimized.

Reference to Table 5 indicates that there are only two instances (and these were both for the sophomore-year male withdrawals) where the correlations were higher than .25. Although the correlation analysis indicates that for the males in Group 2C there is a statistically significant relationship, albeit negative, between academic aptitude and change in stereotypic beliefs and a significant positive relationship between academic aptitude and change in dogmatism,[3] the magnitude of the correlations strongly suggest that the relationship is not of any practical significance. In fact, one might go so far as to say that there is no real relationship between general academic aptitude and changes in stereotypic beliefs, dogmatism, and traditional-value orientation for any of the groups. These findings support those of Plant (1962) who also reported no significant relationship between academic aptitude and changes in dogmatism and ethnocentrism.

It is readily evident from Table 2 that regardless of sex and amount of college education, all groups tended to become less stereotypic in their beliefs, less dogmatic and more receptive to new ideas, and (with the exception of the senior males and females, and junior-year female withdrawals) more traditional-value oriented from 1958–1962. It is further evident from looking at the net changes between 1958 and 1962 that there is no pattern of change. In some groups the males changed more than the females, and vice versa. In addition, there was no pattern evident from one test to another. From the data presented in Table 6, taken in conjunction with that in Table 2, the following conclusions appear to be warranted:

1. There was no significant difference between males and females in the three control groups(freshman, sophomore, and junior-year withdrawals) in regard to the decrease in their stereotypic beliefs and dogmatism between 1958 and 1962.

---

[3] Actually, the changes are in the same direction. The manner of scoring is reversed for the Inventory of Beliefs and Dogmatism Scale so that a high score on one corresponds to a low on the other.

2. Female seniors changed significantly more in their receptivity to new ideas (lessened dogmatism) than did their male counterparts between 1958 and 1962.

3. Both male and female seniors became more emergent-value oriented between 1958 and 1962. However, the female seniors changed to a significantly greater degree than did their male counterparts. It is indeed difficult to explain this phenomenon. However, when one recalls that the females stated they were influenced to a greater degree than males by such experiences as "person dated, close friend/s, and being away from home [Lehmann & Dressel, 1963, p. 133]," it is not unexpected that the influence of the peers and student-subculture matrix resulted in females becoming more outer- or other-directed than their male counterparts.

Whereas the male junior-year withdrawals (3C) became more traditional-value oriented between 1958 and 1962, their female counterparts became more emergent-value oriented. Once again, the difference in amount of change in value orientation between the sexes is significant, especially because the sexes changed in opposite directions. This finding is difficult to interpret inasmuch as both males and females were at college for anywhere from 7–10 terms. Again, the analysis of experiences perceived by these subjects (Lehmann & Dressel, 1963) as having had an impact upon their attitudes and values is helpful. While the males felt that both academic and nonacademic experiences influenced their behavior, the females stated that the nonacademic experiences were most influential. Furthermore, the five most important nonacademic experiences mentioned by females related primarily to the student subculture, and this culture matrix, especially for upperclassmen, is peer oriented. This being the case, it is to be expected that the females would tend to become more emergent or outer-directed in their value orientation than would their male counterparts.

4. Although both male and female freshman-year withdrawals (1C) became more traditional-value oriented between 1958 and 1962, the males changed to a greater degree than did the females.

5. Both male and female seniors became less stereotypic in their beliefs, less authoritarian, and more flexible between their freshman and senior years. Although there was some sex difference when these subjects were tested as freshmen in 1958, there was a marked difference in the amount of change that occurred between the sexes from 1958–1962. During this 4-year period, the females changed significantly more than the males. It is conceivable, of course, that the nonacademic experiences so commonly mentioned by females resulted in their having more contacts (than did their male counterparts) with people differing in race, creed, color, personality, and outlook on life. This being the case, we would have expected the females to change to a greater degree than the males because of a larger universe of experiences.

CONCLUSIONS

The magnitude of change in certain beliefs, ideals, interests, and attitudes certainly leaves much to be desired in terms of the professed aims and objectives of colleges and universities. College personnel should not feel complacent about what has occurred. In fact, one might have expected that if college does have an impact upon personality development, the changes for the seniors would have been consistently more pronounced than those occurring in the withdrawal subjects.

Those who feel that personality modification through college experiences will only be successful with the average or above average in intelligence should realize that this is not necessarily the case. Although the data suggested that the females were more susceptible to change than were their male counterparts, this need not imply that male attitudes and values are fixed and not malleable. The various male groups did change significantly, but those experiences that had an impact upon them differed from those that influenced female personality development. This might suggest, then, that college administrators attempt to capitalize on the experiences unique to each sex rather than attempt to apply the same prescription to all.

Although attitudes and values are instilled early in life and are most easily modifiable in infancy and adolescence, it is readily evident that changes do take place from ages 18–22, or older. It is therefore imperative that our colleges and universities recognize these facts and discard the notion that behavior characteristics are not their concern because it is too late to do anything about them. Just as the myth that intellectual development ceases after a certain age has been discarded, so should the notion that personality change is not possible for college students be discounted. However, a college education, per se, is not instrumental in bringing about these changes, although college attendance might facilitate this development. Therefore, college faculties and administrators must realize that they are not necessarily providing a unique experience for their students, but that maturation and the social environment might have more impact upon personality development than courses and formal academic experiences.

REFERENCES

AMERICAN COUNCIL ON EDUCATION. *Inventory of Beliefs: Instructor's manual.* Washington, D. C.: ACE, 1950.

BARTON, A. H. *Studying the effects of a college education: A methodological examination of changing values in college.* New Haven, Conn.: Edward W. Hazen Foundation, 1959.

BROWN, D. R., & BYSTRYM, DENISE. College environment, personality, and social ideology of three ethic groups. *Journal of Social Psychology,* 1956, **44,** 279–288.

CARLSON, E. R. Attitude change through modification of attitude structure. *Journal of Abnormal and Social Psychology,* 1956, **52,** 256–261.

COREY, S. M. Attitude differences between college classes: A summary and criticism. *Journal of Educational Psychology,* 1936, **27,** 321–330.

EDDY, E. D., JR. Changing values and attitudes on the campus. In, *Long range planning for education.* Washington, D. C.: American Council on Education, 1957. Pp. 137–150.

EDDY, E. D., JR. *The college influence on student character.* Washington, D. C.: American Council on Education, 1959.

HALL, R. M. Religious beliefs and social values. Unpublished doctoral dissertation, Syracuse University, 1950.

HELSON, H., BLAKE, R. R., MOUTON, JANE S., & OLMSTEAD, J. A. Attitudes as adjustments to stimulus, background, and residual factors. *Journal of Abnormal and Social Psychology,* 1956, **52,** 314–322.

JACOB, P. E. *Changing values in college: An exploratory study of the impact of college teaching.* New York: Harper, 1957.

LAZURE, MARTHA C. An intercultural study of personality development in college women of the United States and French Canada. Unpublished master's thesis, Bryn Mawr College, 1959.

LEHMANN, I. J. Changes in critical thinking, attitudes, and values from freshman to senior year. *Journal of Educational Psychology,* 1963, **54,** 305–315.

LEHMANN, I. J., & DRESSEL, P. L. *Critical thinking, attitudes, and values in higher education.* East Lansing: Michigan State Univer. Press, 1962.

LEHMANN, I. J., & DRESSEL, P. L. *Changes in critical thinking, attitudes, and values* *associated with college attendance*. East Lansing: Michigan State Univer. Press, 1963.

LEHMANN, I. J., & PAYNE, ISABELLE, K. An exploration of attitude and value changes of college freshmen. *Personnel and Guidance Journal,* 1963, **41,** 403–408.

MCGUIGAN, F. J. Psychological changes related to intercultural experiences. *Psychological Reports,* 1958, **4,** 55–60.

NEWCOMB, T. M. *Personality and social change*. New York: Dryden, 1943.

NEWCOMB, T. M. Student peer-group influence. In Nevitt Sanford (Ed.), *The American college: A psychological and social interpretation of the higher learning*. New York: Wiley, 1962. Pp. 469–488.

PLANT, W. T. Personality changes associated with a college education. Final report, United States Office of Education, Cooperative Research Branch Project No. 348, 1962.

PRINCE, R. A study of the relationship between individual values and administrative effectiveness in the school situation. Unpublished doctoral dissertation, University of Chicago, 1957.

ROKEACH, M. *The open mind*. New York: Basic Books, 1960.

ROSENBERG, M. Psychological depression and educational attitudes. *Student Medicine,* 1956, **5,** 5–20.

SCHEFFÉ, H. *The analysis of variance*. New York: Wiley, 1959.

SMITH, H. P. Do intercultural experiences affect attitudes? *Journal of Abnormal and Social Psychology,* 1955, **51,** 469–477.

WEBSTER, H. Changes in attitudes during college. *Journal of Educational Psychology,* 1958, **49,** 109–117.

WEBSTER, H., FREEDMAN, M., & HEIST, P. Personality changes in college students. In Nevitt Sanford (Ed.), *The American college: A psychological and social interpretation of the higher learning*. New York: Wiley, 1962. Pp. 811–846.

(Received February 1, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 2, 99-101

OVERT VERSUS COVERT RESPONDING AND RETENTION BY SIXTH-GRADE STUDENTS

ROBERT L. CRIST[1]
Illinois State University

A 6th-grade classroom (N = 24) was randomly divided into 2 groups. Both groups read a programed text, but 1 group responded overtly and the other covertly. Ss were tested immediately upon completion of each programed text and also 6 wks. later. On the immediate posttests there were slight but nonsignificant differences between the overt- and covert-responding groups. On the delayed posttest the slight differences vanished, and the overt and covert groups had almost identical scores. Time data clearly favored the covert group. Comparable replicative data were obtained with a 2nd classroom (N = 20).

Among older students and adults, comparisons of overt and covert responding to programed texts tend to favor covert responding because of the savings in time (Cumming & Goldstein, 1964; Feldhusen, 1963). This study seeks to determine whether similar results will be obtained with elementary-level students.

METHOD

Material

Two programed texts from the Coronet Film Company were used: *Our Solar System* (SS) and *Latitude and Longitude* (LL). They contained 331 and 351 frames, respectively. All tests were congruent with the programs and measured primarily their factual content. Tests over a given program were given immediately upon completion of the program and 6 weeks after the last student had finished the programs.

Subjects

Subjects (*Ss*) were 44 children in the sixth grade of a university laboratory school. Two classrooms were used. One classroom had a mean IQ of 120 (SD = 13.8), and the second classroom (*N* = 24) had a mean IQ of 113 (SD = 15.6).

In both classrooms two groups were formed through randomization.

Procedure

The design is shown in Table 1. It is the same for both classrooms with one exception. Classroom 1 read SS and then LL, while Classroom 2 read LL and then SS.

Ss read the programs 20 minutes each day. Upon completion of a program, each *S* immediately took the test covering the program's factual content.

Ss read only the programed texts, and no supplementary learning experiences were presented. Classroom teachers were always in the room, but they were unobtrusive and provided assistance only when it was requested.

RESULTS

Table 1 shows the group mean achievement scores and mean completion time for each of the programed texts. Standard deviations of all measures are also shown. The data for each classroom are treated separately.

Classroom 1

Table 1 shows that both groups read SS first. The covert-responding group, Group A, had a mean of 85%, while Group B, the overt-responding group, Group B, had a mean of 77%. This difference was not significant at the .05 level. Upon completion of SS, groups changed response mode and read LL. With this program, Group A responded overtly and had a mean of

---

[1] The author wishes to acknowledge and thank Anita Jones and Mary Rozum of the Metcalf School for their cooperation and assistance.

TABLE 1

SUMMARY OF MEAN TEST SCORES AND MEAN TIME FOR COMPLETION OF TWO PROGRAMED TEXTS

Tests	Classroom 1 (*N* = 24)					Classroom 2 (*N* = 20)				
	Group	Solar system		Latitude & longitude		Group	Latitude & longitude		Solar system	
		M	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*
Immediate posttest	A	85[a]	19[a]	86	16	C	87[a]	12[a]	90	20
	B	77	26	79[a]	13[a]	D	86	13	85[a]	13[a]
Delayed posttest	A	83[a]	16[a]	80	15	C	88[a]	21[a]	81	10
	B	83	19	77[a]	11[a]	D	86	13	80[a]	10[a]
Mean time for program completion	A	133[a]	58[a]	162	28	C	193[a]	29[a]	178	33
	B	218	53	137[a]	42[a]	D	273	59	93[a]	14[a]

[a] Covert responding group data.

86%, while Group B responded covertly and had a mean of 79%. This difference was not significant at the .05 level.

Classroom 2

Group C achieved a score of 87% on LL, while Group D had a score of 86%. When response mode was reversed and SS was studied, differences increased with Group C having a score of 90% and Group D, a score of 85%. Neither of the above differences was significant at the .05 level.

Delayed Posttest

The second row of data in Table 1 shows mean scores on the delayed posttest. Although there were slight but nonsignificant differences on the immediate posttest, group differences became even smaller on the delayed posttest.

In Classroom 1, Groups A and B had identical scores of 83% on the test covering SS. On the LL items, Group A had a score of 80 and Group B, a score of 77.

In Classroom 2, differences between groups were equally small. On LL, Group C had a score of 88 and Group D, a score of 86. On the test covering SS, Group C scored 81, and Group D scored 80.

None of the group differences on the delayed posttest approached significance.

Time Data

The third row in Table 1 shows the mean program-completion time. In every instance covert responding took less time than overt responding. This time difference takes its most extreme form in Classroom 2 with SS when the covert-group mean was 93 minutes, while the overt mean was 178 minutes. The large SDs provide evidence of the appreciable differences in the time it took *S*s to complete the programs.

CONCLUSIONS

Differences in retention scores for groups responding overtly and covertly were compared. On both the immediate posttests and the delayed posttests, no significant differences between the overt and covert groups were obtained.

The absence of significant differences in retention coupled with the appreciable saving in time of the covert group provide support for covert responding at the sixth-grade level. Replicative data provided by a second classroom provide additional support for these conclusions.

REFERENCES

CUMMINGS, A., & GOLDSTEIN, L. S. The effect of overt and covert responding on two kinds of learning tasks. In J. P. DeCecco (Ed.), *Educational technology: Readings in programed instruction.* New York: Holt, Rinehart & Winston, 1964. Pp. 231–241.

FELDHUSEN, J. F. Taps for teaching machines. *Phi Delta Kappan,* 1963, **44,** 265–267.

(Received May 10, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 2, 102–108

SEQUENCE OF VOCAL RESPONSE-DIFFERENTIATION TRAINING AND AUDITORY STIMULUS-DISCRIMINATION TRAINING IN BEGINNING FRENCH[1]

LAWRENCE MACE
University of California, Los Angeles

Sequence of speaking and listening in beginning French was studied with 108 early elementary school children. Instruction was presented by an automatic device in 10 daily 15-min. lessons according to 4 treatments: (a) 5 speaking lessons followed by 5 listening lessons, (b) the 5 listening lessons followed by the 5 speaking lessons, (c) 10 concurrent lessons alternating speaking and listening with speaking 1st, (d) concurrent training as in (c), but with listening 1st. At the end of instruction, and 2 wks. following instruction, the speaking-1st groups reliably excelled the listening-1st groups on a listening-comprehension test.

The question behind this research is the role and importance of vocal response learning in the acquisition of language. Specifically, the study investigates the optimal sequence of speaking training and listening training in the teaching of French as a second language. That is, should a person beginning to learn French be taught to speak at the same time he is taught to listen, or should speaking be postponed until after considerable competence in listening has been established?

In this experiment, vocal response learning occurred through a process of *response differentiation* as subjects (*S*s) vocalized different French words and phrases. This learning through speaking was contrasted with learning through listening which occurred without overt vocal responding; listening in the experiment required only *discrimination* among French utterances as auditory stimuli. Three alternative possibilities were explored in order to investigate optimal sequence of vocal response-differentiation training and auditory stimulus-discrimination training: (*a*) Vocal training should be deferred until after the learner has had fairly extensive training in which French utterances, as auditory stimuli, have acquired considerable discriminative control; (*b*) vocal training should occur early, receiving special emphasis *before* extensive auditory discrimination training; and (*c*) concurrent attention should be given both to vocal response and auditory discrimination training.

Nelson Brooks (1960), a leading proponent of the audio-lingual approach to language teaching, maintains that listening (discrimination training) must precede speaking (sound production or vocal response differentiation). He states:

> The learner's activities may be briefly summarized as follows: . . . He is to hear much more than he speaks, he is to speak only on the basis of what he has heard. . . [p. 50]. It is the ear that dominates the learning and use of speech sound. Ear training must come first. . . . The second objective is the reproduction by the tongue and adjacent organs

---

[1] This research was performed under a contract with the United States Office of Education, Department of Health, Education, and Welfare (NDEA, Title VI, Contract No. OE-3-14-022, Evan R. Keislar, principal investigator). Appreciation is expressed to the principal and staff of the Nora Sterry Elementary School for their cooperation.

of the speech sounds the ear has learned to recognize [p. 107].

Carroll (1963) and Lane (1965b), in two critical reviews of the literature on research in second-language learning, each report a lack of agreement among language specialists and a dearth of experimental evidence on the priority of either discrimination training or vocal response training in language learning. However, the view that discrimination training should take priority over sound-production training has been accepted almost as dogma by many language teachers. This view also has been followed by a number of specialists in programed self-instruction as a basis for preparing foreign-language programs (cf. Morton, 1960; Sapon, 1961).

A role and possible importance of speech production during language learning is suggested by a series of investigations in speech phonology by Liberman (1957) and his co-workers who advance a *motor theory of speech perception*. Each phoneme of a language comprises a range or category of physically different sounds which are classified as one sound by the members of a language community. Liberman proposes that the individual, as he listens to speech, emits sounds falling within the phoneme categories, and that these vocal or subvocal responses mediate speech-sound perception. This view implies that training in speech-sound production may be important in learning to hear the phonemes of a language.

Children learning French in the present experiment were taught not only to articulate and to distinguish French phonemes and phoneme combinations, but also to *comprehend* these sound patterns. Speaking and listening comprehension were defined according to the concepts of response differentiation and stimulus discrimination. Response differentiation occurred during speaking training as *S*s learned to articulate different French words and phrases; stimulus discrimination occurred also during speaking training as pictures were used to evoke these utterances. Consequently, through speaking training, pictures became discriminative stimuli controlling newly differentiated vocal responses. A child in the experiment spoke with comprehension if he made appropriate utterances when particular pictures were presented. During listening training, stimulus discrimination was the only process which was emphasized: A French utterance was presented as an auditory stimulus, and *S*s learned to select a certain picture from among several alternative pictures. The extent to which a child learned to select pictures corresponding to particular utterances indicated his listening comprehension.

Speaking training required both the differentiation of new vocal responses and also the stimulus control of these responses through discrimination; listening training emphasized only discrimination. Sequence of vocal response differentiation and auditory stimulus discrimination was investigated by arranging materials for experimental treatments with either speaking training first or listening training first. The same materials were also rearranged to provide treatments for concurrent speaking and listening training.

METHOD

Subjects

*S*s were 108 children from six classrooms: 41 first graders, 43 second graders, and 24 third graders. Many *S*s were from Japanese homes (56%) or Mexican-American homes (21%). However, all *S*s were judged by their teachers to be capable of following instructions in English. None of the *S*s had a

French language background. The school is located in a lower-middle-class socioeconomic neighborhood in Los Angeles, California.

Equipment

For presentation of stimuli, a Wollensak tape recorder was synchronized with a 35-millimeter Kodak Carousel slide projector by means of a Kodak Cavalcade programer. Groups of four to six *S*s sat at tables separated by plywood partitions. They listened to taped audio materials through earphones and watched visual presentations which were projected on a screen before them. The plywood partitions and the earphones helped to limit social interaction among *S*s.

Materials and Procedures

A 10-day 600-frame program of beginning French instruction consisted of five daily lessons (300 frames) of speaking training, and five lessons (300 frames) of listening training. Each lesson, 60 frames in length, lasted 10–12 minutes. Through the duplication and rearrangement of the 600 frames, four sequences of materials were created, one for each of the four experimental groups described below. These sequences consisted of identical frames arranged in different orders.

For each of the 300 speaking frames in the program, a single picture was projected upon the screen. *S*s responded by making an utterance. Then the correct utterance was presented from the tape recorder. The correct utterance depended upon the lesson being taught. For example, in the early lessons of the program a picture of a large dog in a standing position might require the utterance, "*un chien*" (a dog), while later in the program the required response to the same picture might be "*Le grand chien est debout.*" (The large dog is standing.)

For each of the 300 listening frames, three numbered pictures were projected on the screen simultaneously. For example, the pictures for a frame might be: a large girl in a standing position (1), a small dog in a standing position (2), and a small girl in a sitting position (3). A French utterance such as "*La petite fille est assise*" (The small girl is sitting.) was presented from the tape recorder, and *S*s selected one of the three pictures on the screen, writing its corresponding number on a scoring sheet. Confirmation was provided as the number of the correct picture (in this example, the number 3) was spoken in English from the tape recorder.

*S*s were assigned randomly to four treatment groups, 27 in each group. The first group received massed training in speaking followed by massed training in listening; five daily lessons of speaking training were followed by five daily lessons of listening training. The second group received massed training in listening followed by massed training in speaking; five daily lessons of listening training were followed by five daily lessons of speaking training. The third and fourth groups received the two types of training concurrently; that is, each of the 10 daily lessons included both kinds of training. Each day for 10 days the third group received 15-frame segments of speaking training followed by 15-frame segments of listening training involving the same utterances. The training for the fourth group was the same as for the third group, except that each day the listening-training segments came *before* their speaking counterparts. To summarize, therefore, the first and third groups received speaking training first, while the second and fourth groups received listening training first; the first and second groups received massed training in speaking and in listening, while the second and third groups received approximately concurrent training in speaking and listening.

Following the four instructional treatments, a Listening Comprehension Test was administered to all *S*s in groups of six on one school day. The 60 items of this test covered the entire content of the program and were similar to the frames of the listening training with the exceptions that no confirmation was provided, and 23 of the test items included six numbered alternative pictures rather than three. Each *S*'s score on this Listening Comprehension Test was the total number of items answered correctly. This same Listening Comprehension Test was administered a second time 2 weeks following instruction in order to measure retention.

During the 2 weeks following instruction, each *S* was given an individual speaking test with 80 items similar to the frames of the speaking portions of the instructional program except that no confirmation was provided. For each item of the Speaking Test a picture was projected on the screen, and *S* spoke into a microphone recording his utterance. A native French judge scored each *S*'s speaking performance, allowing points for the appropriateness and completeness of *S*'s utterance as a response to the picture. An *S*'s score on the Speaking Test was the total points achieved from a possible 324 points.

TABLE 1
ADJUSTED TREATMENT MEANS AND STANDARD DEVIATIONS FOR SCORES ON TESTS OF LISTENING COMPREHENSION, RETENTION OF LISTENING COMPREHENSION, AND SPEAKING ACHIEVEMENT

Treatments	Listening comprehension		Retention of listening comprehension		Speaking achievement	
	M	*SD*	*M*	*SD*	*M*	*SD*
5 speaking lessons before 5 listening lessons	45.73	13.78	49.35	12.22	152.75	79.34
5 listening lessons before 5 speaking lessons	36.25	16.28	36.64	16.97	138.79	76.35
Speaking before listening in the same lessons	41.03	13.48	43.87	12.30	114.32	83.31
Listening before speaking in the same lessons	39.78	16.44	39.85	16.62	137.47	59.64

RESULTS

A correlation of .92 was obtained between scores on the Listening Comprehension Test administered at the end of the instructional program and the scores on the same test administered 2 weeks following instruction; this correlation suggested that the retest reliability of the Listening Comprehension Test was satisfactory. The split-half reliability of the scores on the Speaking Test was found to be .97, indicating that the Speaking Test was also reliable.

Reliable correlations of .49 or higher were noted between *Ss*' mental ages and their scores on the three posttests. Consequently, the means for each of the four treatment groups on each posttest were adjusted for mental age. Table 1 presents the adjusted means and standard deviations for each of the treatment groups on each posttest.

Analyses of covariance were performed on the scores of the three posttests using mental age as a covariate. To study the effect of sequence on each of the three posttests, the scores of the speaking-first groups were compared with the scores of the listening-first groups. A statistically significant difference ($F = 4.56$, $df = 1/103$, $p < .05$) was found for the effect of sequence on the Listening Comprehension Test administered at the end of instruction. This difference favored the speaking-first groups, suggesting that the speaking-first treatments were more effective than the listening-first treatments in producing performance in listening comprehension immediately following the program of instruction.

Also, a statistically significant difference ($F = 11.95$, $df = 1/103$, $p < .01$) favored the speaking-first groups when the Listening Comprehension Test was administered 2 weeks following instruction. This finding suggests that the speaking-first treatments produced superior retention of performance in listening comprehension.

The effect of massed training, beginning either with 1 week of speaking training or 1 week of listening training, was compared with the effect of concurrent daily training in speaking and listening. No statistically significant differences were found for either the Listening Comprehension Test or the Retention Test between scores of the two groups which had massed training and scores of the two groups which had concurrent training. Also, no statistically significant differences were detected on the Speaking Test for either the effect of sequence or the effect of massed versus concurrent training.

DISCUSSION

The results indicate that *sequence* of speaking and listening training was important in the experiment: The speaking-first treatment groups performed better than did the listening-first groups on both administrations of the Listening Comprehension Test.

One explanation for this finding is that speaking training produced some degree of listening proficiency, and, consequently, the groups which received speaking training first may have begun listening training with some part of the listening task already learned. This advantage may have resulted from (*a*) increased discriminability of stimuli due to articulatory and verbal labeling, and (*b*) reversibility of stimulus and response terms.

Articulatory and Verbal Labeling

Articulatory responses were taught in the experiment during speaking training: A French utterance was presented as an auditory stimulus, and each *S* responded by voicing the utterance himself. This training may have facilitated auditory discrimination for the speaking-first groups during subsequent listening training. Liberman (1957) proposes that an articulatory response may be learned as a label for a range or class of nonidentical acoustic stimuli, and that distinctive proprioceptive stimuli from the articulatory response may add to the stimuli of a class of speech sounds making them more distinctive or discriminable. Holland and Matthews (as cited by Lane, 1965a) also suggest that vocalization may play an important role in speech discrimination by supplying supplemental kinesthetic stimuli. The present results support this view that learned articulatory responses facilitate subsequent auditory discrimination.

Considerable research suggests that *verbal* labeling of stimuli may increase the discriminative control of these same stimuli under other circumstances (cf. Albert, 1959; Goss & Greenfield, 1958; Norcross & Spiker, 1957; Reese, 1960). Through speaking training in the present experiment, different French utterances became verbal labels for certain visual stimuli. For example, *S* learned to say "*La fille est assise*" when a picture of a girl sitting was presented, and "*La fille est debout*" when a picture was that of a girl standing. In this example, the sitting-standing aspect of the picture became a controlling visual stimulus for the vocal response. This training could have been useful to the speaking-first *S*s during listening training when one of these utterances was presented, and *S*s were required to select a picture from among alternatives on the basis of a particular visual property.

Reversibility of Stimulus and Response Terms

Research in paired-associate learning has demonstrated that when a response, *B*, is learned to a stimulus, *A*, a reversibility of these stimulus and response terms occurs so that *B* as a stimulus is likely to evoke *A* as a response. This effect has been described as *backward* or *reverse asso-*

ciation in which learning $A \rightarrow B$ simultaneously establishes a $B \rightarrow A$ connection (cf. Cramer & Cofer, 1960; Feldman & Underwood, 1957; Kjeldergaard & Horton, 1960; Mace & Keislar, 1965; McCormack, 1961; Murdock, 1956). Although stimulus-response reversibility itself needs explanation, this effect may be used to account for the present findings. During speaking training, the speaking-first groups learned utterances as vocal responses to pictures as visual stimuli. Subsequent listening training for the speaking-first groups required a reversal of these stimulus and response terms; *S*s responded by selecting a picture from alternatives when the utterance was presented as an auditory stimulus. Through reversibility during speaking training the speaking-first groups may have acquired listening competence prior to listening training. Initial listening competence could have led the speaking-first groups to superior listening performance.

The results, accounted for by labeling and reversibility, suggest that the most effective sequence for teaching listening comprehension is speaking training before listening training. More generally, the results suggest that early vocal response differentiation may be important in language learning. The experiment is not conclusive, however, since two extraneous explanations for the findings are plausible. These are: (*a*) possible differences in group morale, and (*b*) recency of discrimination training.

Speaking-First Group Morale

During speaking training, six *S*s spoke aloud as a picture was presented on the screen; social interaction among *S*s at this time may have encouraged competition and high morale. Higher morale for the speaking-first groups may have carried over during subsequent listening training leading to superior listening performance. In order to rule out this interpretation, it would have been necessary to administer treatments individually to *S*s.

Recency of Discrimination Training

When the Listening Comprehension Test was administered at the conclusion of instruction, one speaking-first group had just completed a full week of listening training. Therefore, this group may have had less time to "forget" before the Listening Comprehension Test than the listening-first group which had completed listening training a week earlier. Although a difference between the groups held up well 2 weeks following instruction when the Listening Comprehension Test was administered a second time, the effect of recency of training cannot be discounted. No control appears completely satisfactory for obviating this recency interpretation. However, if all testing had been postponed for several weeks or more following instruction and the same findings had prevailed, then an explanation in terms of recency would be less credible.

Differences among the groups on the Speaking Test would have helped greatly in finding the optimal sequence of speaking and listening training. If the speaking-first groups had excelled on both the Listening Comprehension Test and the Speaking Test, the recency explanation could be discounted, and speaking first could be recommended generally. On the other hand, if the listening-first groups had excelled on the Speaking Test, while the speaking-first groups excelled on the Listening Comprehension Test, then the effect of recency would be clearly suggested. A more sensitive Speaking Test was needed in the experiment.

In conclusion, the experiment gives no support to the formal tenet of the audio-lingual approach to language teaching (Brooks, 1960, p. 107) which recommends that training in listening comprehension should precede sound-production training. In fact, in terms of the results obtained, the experiment would support the opposite conclusion; that is, for most effective acquisition of listening-comprehension skills, differentiation of utterances as vocal responses through speaking should precede training which emphasizes only auditory discrimination.

REFERENCES

ALBERT, R. S. Function of verbal labels in the discrimination of subtle stimulus differences. *Journal of Genetic Psychology,* 1959, **94,** 287–296.

BROOKS, N. *Language and language learning: Theory and practice.* New York: Harcourt, Brace & World, 1960.

CARROLL, J. B. Research on teaching foreign languages. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963.

CRAMER, PHOEBE, & COFER, C. N. The role of forward and reverse association in transfer of training. *American Psychologist,* 1960, **15,** 463.

FELDMAN, S. M., & UNDERWOOD, B. J. Stimulus recall following paired-associate learning. *Journal of Experimental Psychology,* 1957, **53,** 11–15.

GOSS, A. E., & GREENFIELD, N. Transfer of a motor task as influenced by conditions and degree of prior discrimination training. *Journal of Experimental Psychology,* 1958, **55,** 258–269.

KJELDERGAARD, P. M., & HORTON, D. L. An experimental analysis of associative factors in stimulus equivalence, response equivalence, and chaining paradigms. *Studies in Verbal Behavior.* Report No. 3. Minneapolis: Univer. Minnesota Press, 1960.

LANE, H. Programmed learning of a second language. *International Review of Applied Linguistics,* 1965, **2,** 249–301. (a)

LANE, H. Programmed learning of a second language. In R. Glaser (Ed.), *Teaching machines and programmed learning: Data and directions.* Washington, D. C.: National Education Association, 1965. Pp. 584–643. (b)

LIBERMAN, A. M. Some results of research on speech perception. *Journal of the Acoustical Society of America,* 1957, **29,** 117–123.

MACE, L. L., & KEISLAR, E. R. Reversibility of stimulus and response terms following discrimination learning of French phonemes. *Journal of Educational Psychology,* 1965, **56,** 46–49.

MCCORMACK, P. D. Backward mediated positive transfer in a paired-associate task. *Journal of Experimental Psychology,* 1961, **61,** 138–141.

MORTON, F. R. The language laboratory as a teaching machine. Report No. 1, Language Laboratory, University of Michigan, 1960.

MURDOCK, B. B. Backward learning in paired associates. *Journal of Experimental Psychology,* 1956, **51,** 213–215.

NORCROSS, KATHRYN, & SPIKER, C. C. Effects of type stimulus pretraining on discrimination pretraining performance in preschool children. *Child Development,* 1957, **28,** 80–84.

REESE, H. W. Motor-paired learning and stimulus pretraining. *Child Development,* 1960, **31,** 505–515.

SAPON, S. M. Report on the use of a programmed course in Spanish. Unpublished report, Ohio State University, 1961.

(Received July 2, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 2, 109-114

BLACKOUT RATIO AND OVERT RESPONSES IN PROGRAMED INSTRUCTION:

RESOLUTION OF DISPARATE RESULTS[1]

FREDERICK D. KEMP AND JAMES G. HOLLAND[2]
Harvard University

The blackout measure, for the degree to which material is programed, was applied to 12 sets of material used in previous studies of overt vs. covert responding in teaching-machine programs. Blackout ratios (% total words which could be removed without affecting error rate) ranged from 11.1%-74.6%. The lowest 4 ratios (11.1%-25.4%) were for programs which previously demonstrated an advantage for overt responding. The remaining 8 ratios (31.0%-74.6%) were for programs which previously yielded no response-mode difference. Many studies have failed to find an advantage for overt responding because the material used was largely unprogramed in that answers were not dependent upon much of the content of the material.

The rationale underlying programed instruction demands that there be a response to each item in a program. It has also been suggested that the response should be overt to assure that responses occur without fail. However, the now voluminous research literature contains inconsistent results for comparisons of posttest performance after overt response and after either the reading of completed statements or "thinking" the answer. Many studies have failed to find a statistically significant difference, but others have confirmed the advantage of an overt answer. (For a review of relevant studies see Holland, 1965.) One reason suggested for the frequent failure to obtain a significant difference has been that "poorly" programed or unprogramed material may have been used in many of the studies. This point received some support in the results of an experiment (Holland, 1964) in which a program was deliberately altered by leaving blank for the subjects' (*S*s') completion (*a*) words that could be supplied quite easily, but for reasons relatively unrelated to the principal content of the items, and (*b*) words, which, although relevant to the critical content of the item, could not readily be given by *S*s. Neither of the experimental programs showed significant differences from reading completed statements. However, the normal program, in which the answer term was both relevant to the critical content and was to be provided by *S*, showed a significant advantage for overt responding.

To determine experimentally whether response mode or other characteristics of programed materials are important, it is necessary to have an unequivocal, preferably quantitative, experimental definition of the degree to which the material used is programed. Such a measure has recently been developed (Holland &

---

[1] The research reported in this paper was sponsored by the Air Force Systems Command, Electronic Systems Division, Decision Sciences Laboratory, under Contract No. AF19(628)-2404 with Harvard University. The paper is identified as ESD Technical Documentary Report No. ESD-TDR-65-355. Further reproduction is authorized to satisfy needs of the United States Government.

The research was conducted in the facilities provided by the Harvard Committee on Programmed Instruction which is supported by the Carnegie Corporation.

[2] Now at University of Pittsburgh.

Kemp, 1965). This measure, the blackout ratio, consists of obliterating all material in a program which can be removed without affecting error rate. Such material is unprogramed in the sense that answers are not contingent upon (nor dependent upon) mastery of that material. The quantitative index is expressed as the ratio of the number of words that can be so obliterated to the total number of words in the material. By applying the blackout technique to programs used in previous response-mode studies, this experiment evaluates the claim that the frequent use of relatively unprogramed materials has caused the existing ambiguity in the literature regarding the value of an overt response. Excluding high-error-rate programs from consideration, programs which revealed an advantage for overt responding would be expected to have a lower blackout ratio, on an average, than programs which failed to show a difference between overt and covert responding.

METHOD

Blackout Technique

The blacking-out procedure consists of physically obliterating those portions of an item judged to be unnecessary to the correct response. Generally, only whole phrases are blacked out; articles, prepositions, and auxiliary words are blacked out only with the full phrase. In calculating the blackout percentage, the unit used is, simply, words. Blanks are not counted as words and are never blacked out. Where diagrams, mathematical expressions, etc., appear in an item, they are treated somewhat arbitrarily, but consistently, as to what constitutes a "unit." Diagrams are not blacked out (only the Cardiology Program contained many diagrams); thus, the blackout percentage refers strictly to verbal content. The supplementary materials ("panels") are not blacked out.

Figure 1 presents several samples of the same items from different programs, in normal and blacked-out forms. The range of the blackout ratio is apparent.

After blacking out the material (covering it with black crayon), the blacked-out and un-blacked-out versions are tested with two groups of *S*s to validate the assumptions that answers are independent of the material obliterated. When error rate has been significantly influenced, the blackout must be revised, and the new blackout tested until error rate is unaffected. In practice, the first blackout is often adequate, and there is seldom more than one revision required.

Materials and Apparatus

Table 1 lists the programs used and the abbreviations by which they will be referred below. The Cardiology Program is listed on two rows because the Cummings and Goldstein (1962) study classified the items into those with verbal answers and those with pictorial. They had expected that overt responding would be important for pictorial but not for verbal items. However, they found an advantage for overt responding in both types of items.

To keep the size of the experiment within manageable and economical limits, the blackout ratio was determined usually for the first 50 items, but special considerations led to special exceptions. Only the 17 items in the appendix of a report (Roe, Massey, Weltman, & Leeds, 1960) were available for the Statistics Program. The full Cardiology Program was only slightly longer than the 100 items needed to obtain 50 pictorial items; 375 items were used for each of the two forms of the Analysis of Behavior because of the need for this information in relation to another experiment (Holland, 1964). Five programs—AB-A, AB-B, IBM, MG, and Elect (Z)—used supplementary material in the form of panels, which were given to *S* at the appropriate times.

All programs, in either normal or blacked-out forms, were used in teaching machines in the Harvard Self-Instruction Room under supervision of a trained attendant. The AB-A and the AB-B programs were used in the Harvard disc machine; all others were used in the Rheem Didak #501 Machine. Some of the cardiology frames were too large for this machine and required a large answer space. For these items, a special booklet was arranged.

All blackouts in the present study were performed by the senior author. An additional experiment served as a check on his objectivity and on the objectivity of the method itself. Ten naïve *S*s were trained using a teaching-machine program on the blackout technique and compared with one

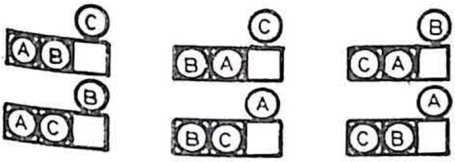
Normal	Blacked out
Stat This illustration shows Cells 1 and 2 filled in all the possible ways they can be filled. Cell 1 = 3 ways, Cell 2 = 2 ways.	[redacted]
 C A B C B A B C A B A C A B C A C B	
We can see in this illustration that there is only one way (ball) left to fill cell 3 when cells 1 and 2 are filled. Thus, there are 3 x 2 x 1 = ____ ways in which 3 balls can fill 3 cells.	[redacted] 3 x 2 x 1 = ____ [redacted]
Card-V In the word "electrocardiogram" indicate which parts mean "electrical", "heart", and "record."	In the word "electrocardiogram" indicate which parts mean "electrical", "heart", and "record."
AB-A If a response which has been accidentally reinforced does not happen to occur just before a second operation of the magazine, it will be ____.	If a response which has been accidentally reinforced does not happen to occur just before a second operation of the magazine, it will be ____.
AB-B If a ____ which has been accidentally reinforced does not happen to occur just before a second operation of the magazine, it will be extinguished.	If a ____ which has been accidentally reinforced [redacted]

FIG. 1. Sample items from four programs as they appeared in the normal and blacked-out versions.

another and with the present "blackout specialist" on the blackout ratios for segments of seven programs. The blackout ratios assigned by 9 of the 10 *S*s correlated highly with one another. The highest correlation between naïve *S*s' blackouts was .988. Of the 45 coefficients 26 exceeded .80, and 9 of the remaining involved the deviant *S*. The 9 *S*s whose blackouts correlated with one another also yielded high correlations with the blackouts of the blackout specialist in the present study (coefficients ranging from .614–.923).

Subjects

A total of 195 undergraduates at Harvard University and Radcliffe College served as *S*s. They were paid a flat fee for participating in the experiment, based on a rate of $1.50 an hour for a number of hours estimated to be sufficient for relatively slow workers.

TABLE 1

EXPERIMENTAL PROGRAMS

Program and abbreviation	Author of program	Previous experiment(s) on overt vs. covert responding	No. items	Difference (D) or no difference (ND)
The Diagnosis of Myocardial Infarction—Pictorial Card P	Basic Systems (unpubl.)	Cummings & Goldstein (1962)	57	D
The Diagnosis of Myocardial Infarction—Verbal Card V	Basic Systems (unpubl.)	Cummings & Goldstein (1962)	124	D
Principles of Missile Guidance MG	TMI (Air Force—unpubl.)	Cartier (1963)	100	D
Analysis of Behavior—Version A AB-A	Holland & Skinner (1959)	Holland (1960, 1964) Williams (1963)	375	D
Genetics Genet	Kormondy (unpubl.)[a]	Kormondy & Van Atta (1962)	50	ND
Electricity Elect (Z)	U.S. Navy basic electricity course, Zuckerman, Marshall, & Groesberg (1961)	Zuckerman et al. (1961)	50	ND
Binary Numbers BN	Silverman & Alter (1961)	Silverman & Alter (1961)	50	ND
Experimental Programmed Course: IBM Sorters IBM	IBM (unpubl.)	Hartman, Morrison, & Carlson (1963)	41	ND
Basic Electricity Elect (S)	Silverman & Alter (1961)	Silverman & Alter (1961)	50	ND
Sets, Relations, & Functions Sets	Eigen (1961)	Lambert, Miller, & Wiley (1962)	50	ND
Analysis of Behavior—Version B AB-B	Holland & Skinner (1959)	Holland (1960, 1964)	375	ND
Statistics Stat	Roe et al. (1960)	Roe et al. (1960)	17	ND

[a] This program is considerably different from a published program which was based on it.

Procedure

The basic procedure involved a comparison between 15 *S*s taking the normal version of the program and 15 *S*s taking the blacked-out version. Most of the data were collected in the summer of 1964. At that time, 30 *S*s took AB-A; 30 took AB-B; 30 took Card, Elect (Z), and Sets; and 30 took IBM, Genet, Elect (S), and BN. Earlier, 30 *S*s took Stat (Spring 1964), and 30 took MG in the fall of 1965.

Preliminary data indicated the necessity of revising the blacked-out versions of four programs—Stat, IBM, Elect (Z), and Sets. This was done in the fall of 1965, when 15 *S*s took the revised version of all four programs. The data were then compared with the data from the normal program which had been gathered earlier.

RESULTS AND DISCUSSION

Blackout Validation

Table 2 presents the medians and quartiles of the percentages of correct responses on the various programs for the normal and blacked-out versions of each. Differences in percentage points between normal and blacked-out versions range from 7.4 (Stat) to −3.2 (Card V), but *no* program showed a significant difference in error rate, as measured by the Kolmogorov-Smirnov two-sample test, one tailed (Siegel, 1956).

TABLE 2

PERCENT CORRECT ON EXPERIMENTAL PROGRAMS: MEDIANS AND INTERQUARTILE RANGES FOR NORMAL AND BLACKED-OUT VERSIONS

Program	Normal		Blacked-Out	
	Median	Interquartile range	Median	Interquartile range
Card P	75.4	54.4–86.4	77.2	73.2–86.0
Card V	68.6	60.3–83.3	71.8	68.4–79.8
MG	93.0	92.0–96.5	93.0	87.2–94.2
AB-A	78.0	61.7–88.4	78.4	70.8–82.9
Genet	82.0	77.5–94.0	82.0	68.0–90.5
Elec (Z)	84.0	73.0–94.0	84.0	72.0–91.0
BN	90.0	81.5–96.0	88.0	77.5–92.0
IBM	85.4	81.7–90.2	80.5	67.7–91.5
Elect (S)	82.0	69.5–92.0	76.0	66.0–84.5
Sets	98.0	96.0–100.0	94.0	91.5–98.0
AB-B	93.6	85.3–97.2	91.7	89.5–93.3
Stat	94.1	82.4–95.6	88.2	80.9–95.6

Note.—No *p*'s reached significance.

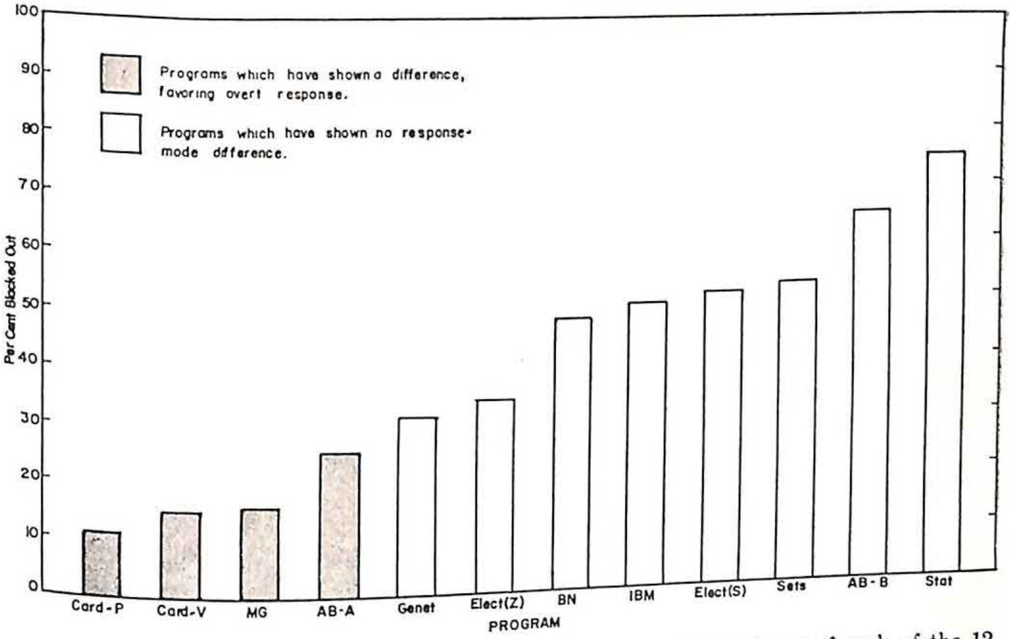


FIG. 2. Percentage of the total number of words which were blacked out of each of the 12 sets of material. (The programs are arranged in rank order from lowest to highest blackout ratios.)

Blackout Ratios

Figure 2 shows the percentage of material blacked out of each of the programs. Those programs which have previously shown a difference favoring overt responding are darkened, while those which have shown no difference are plain. The 12 programs yielded a wide range of blackout ratios (11.1%–74.6%). Aside from the primary purpose of this study, it indicates that the blackout ratio as a measure of program adequacy can vary considerably. In theory the measure can vary from 0%–100%, but in practice the extremes are most unlikely. However, a range of 64 percentage points among 12 programs which were not specifically chosen to explore the range is promising for the blackout measure of programing.

It is readily apparent from Figure 2 that the four sets of material from which the least material could be removed are those which in previous research yielded results favoring an overt response. These had blackout ratios ranging from 11.1–25.4, while the other eight sets of material, for programs previously yielding no difference between overt and covert responding, had blackout ratios ranging from 31.0–74.6. In short, the degree to which material is programed is related to whether or not overt responding is important. If the response is unrelated to the critical content on which *S* is later to be posttested, it makes no difference whether, much less how, he responds, but if *S* is tested on things which have served as a contingency for correct responses, overt responding is important.

REFERENCES

CARTIER, F. A. Comparison of overt and covert responding on a programed lesson assigned as homework. *Journal of Programed Instruction*, 1963, **2**, 13–19.

CUMMINGS, ALLANA Q., & GOLDSTEIN, F. S. The effect of overt and covert responding on two kinds of learning tasks. Technical

Report No. 620919, Center for Programed Instruction, 1962.

EIGEN, L. D. *Sets, relations, and functions.* New York: Center for Programed Instruction, 1961.

HARTMAN, T. F., MORRISON, B. A., & CARLSON, M. E. Active responding in programed learning materials. *Journal of Applied Psychology,* 1963, **47,** 343-347.

HOLLAND, J. G. Design and use of a teaching machine program. Paper read at American Psychological Association, Chicago, September 1960.

HOLLAND, J. G. Response contingencies in teaching-machine programs. *Journal of Programed Instruction,* 1964, **3,** 1-8.

HOLLAND, J. G. Research on programming variables. In R. Glaser (Ed.), *Teaching machines and programmed learning: II. Data and direction.* Washington, D.C.: National Education Association, 1965, Pp. 66-117.

HOLLAND, J. G., & KEMP, F. D. A measure of programing in teaching-machine material. *Journal of Educational Psychology,* 1965, **56,** 264-269.

HOLLAND, J. G., & SKINNER, B. F. A self-tutoring introduction to a science of behavior. Harvard University, 1959. (Mimeo)

KORMONDY, E. J., & VAN ATTA, E. L. Experiment in self-instruction in general biology. *Ohio Journal of Science,* 1962, **4,** 4-10.

LAMBERT, P., MILLER, D. M., & WILEY, D. E. Experimental folklore and experimentation: The study of programmed learning in the Wauwatosa public schools. *Journal of Educational Research,* 1962, **4,** 4-10.

ROE, ANNE, MASSEY, MILDRED, WELTMAN, G., & LEEDS, D. Automated teaching methods using linear programs. Automated Learning Research Project No. 60-105, University of California, 1960.

SIEGEL, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

SILVERMAN, R. E., & ALTER, MILLICENT. Response mode, pacing, and motivational effects in teaching machines. United States Naval Training Device Center, Port Washington, N. Y., Technical Report No. 507-3, 1961.

WILLIAMS, JOANNA P. A comparison of several response modes in a review program. *Journal of Educational Psychology,* 1963, **54,** 253-260.

ZUCKERMAN, C. B., MARSHALL, G. R., & GROESBERG, S. Research in the automation of teaching. United States Naval Training Device Center, Port Washington, N. Y., Technical Report No. 661-1, 1961.

(Received August 2, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 2, 115-120

STRUCTURE OF EXPERIENCE AND FUNCTIONAL FIXEDNESS

GEORGE D. YONGE
Department of Education, University of California, Davis

According to Flavell, functional fixedness varies with the number of unusual uses experienced in the training period. His supporting data are ambiguous owing to a faulty design. The present study focused on his limiting condition(i.e., a switch experienced in its usual and 4 unusual functions), but varied the position (1st, middle, last) of the switch experience. 115 female university students were each assigned to 1 of 5 groups. Contrary to Flavell's hypothesis, if the switch function occurred 1st in the series, functional fixedness was manifested even though 4 unusual uses were subsequently experienced. The structure or gestalt of the training experiences was said to vary depending on the position of the switch function. A phenomenological interpretation of the results was given.

Functional fixedness is usually classified with other "set" phenomena in problem solving and is interpreted as an instance in which previous experience has a negative effect on problem-solving behavior (Birch & Rabinowitz, 1951). For example, if a subject (*S*) uses a pair of pliers as pliers in one situation and is then confronted with a situation in which they must be used as a weight, the previous plier experience will reduce the probability that *S* will use them as a weight if there is another equally appropriate object available which can serve this function. Presumably, the use of the pliers as pliers reinforces this function such that it becomes fixed to the object, and other possible functions are made relatively less available. Jensen (1960) pointed out that the above "negative" interpretation places too much emphasis on the object and not enough on *S*. As Merleau-Ponty (1962) states, there are three terms in any perception: a *figure* on a *ground* for a *subject*, and the perceiver has a privileged role in that what appears as figure or ground is greatly influenced by his purposes.

The meaning of a familiar object stands out as a figure against a background of discoverable meanings. The emphasized and discoverable meanings of an object are correlated with its physiognomy or general shape and style. Very often, with regard to an unfamiliar object, we know what it is *not* before we know what it *is*, and this negative knowledge is based primarily on its physiognomy—its size, shape, weight, etc.[1] The usual two-object functional-fixedness design emphasizes the "figure" (known) meaning[2] of one object, and the ensuing problem situation compares the ease of finding a "ground" (discoverable) meaning of the previously emphasized object with the ease of finding such a meaning in a previously nonutilized object. The control group is used to demonstrate that the two objects have similar discoverable meanings in the problem situation.

---

[1] Husserl's views on typification and relevance, as discussed by Schutz (1962), are germane. However, the extended treatment required to present these ideas is not warranted by the scope of the present paper.

[2] It should be noted that an experience emphasizing a nonessential or unusual use does not hide nor make less discoverable the common or essential function of an object (van de Geer, 1957).

TABLE 1

AMBIGUITY OF THE PREPROBLEM TRAINING ASPECT OF FLAVELL'S DESIGN

Experimental group	Proportion of *S*s experiencing the object in its usual function first	No. unusual uses experienced
E_1	1.00	0
E_2	.50	1
E_3	.33	2
E_4	.25	3
E_5	.20	4

The purpose of the present study was twofold. The first purpose was to determine whether the significance or meaning of several experiences with an object varied depending upon the total structure (gestalt) of these experiences.

The second purpose was to test a consequence of Flavell, Cooper, and Loiselle's (1958) hypothesis that there is a progressive decrease in functional fixedness as the number of novel, unusual functions experienced in the preproblem training period is increased. Although Flavell claimed experimental support for this hypothesis, a flaw in his design renders his data ambiguous. Specifically, it is not known whether his results reflect the *number* of novel functions experienced or the *position* of the usual function in the series of preproblem experiences (i.e., the gestalt of preproblem experiences). It can be seen from Table 1 that there is an inverse relationship between the number of unusual uses experienced and the proportion of *S*s in a given experimental group who experienced the usual function first. For example, one-half of the *S*s in Flavell's E_2 group experienced the usual function first (along with *one* unusual function), whereas only one-fifth of the *S*s in his E_5 group experienced the usual function first (along with *four* unusual functions).

In the present study the position of the usual function in the preproblem series of experiences was varied while keeping the number of unusual uses constant (four—as in Flavell's limiting condition). If Flavell's hypothesis is correct, the position of the usual function will have no effect, and the experimental groups will not differ in the incidence of functional fixedness. Furthermore, the experimental groups will not differ from the control group; that is, there will be no functional-fixedness effect. If the position of the usual function is important, the experimental groups will differ, and this will call into question the validity of Flavell's conclusions.

METHOD

Subjects

All *S*s were female students enrolled in an introductory educational psychology course at the University of California (Davis). The majority were juniors and seniors intending to obtain a teaching credential. Each of the 115 *S*s was assigned to one of five groups in order of appearance at the experimental room.

Experimental Groups

The control group's *S*s (C), without prior experimental training, were given a variation of the Maier (1933) two-string problem which could be solved only by using either a microswitch or a small relay as a weight to tie to one of the strings. This group served to demonstrate the functional equivalence of the two available objects for the two-string problem and to provide the base line by which the presence of functional fixedness was assessed.

*S*s in Experimental Group 1 (E_1), the conventional experimental group in a two-object functional-fixedness paradigm, experienced the switch as a switch and were then confronted with the two-string problem. In the preproblem training period, the members of groups E_2, E_3, and E_4 all experienced the switch in its usual and in four unusual functions. *S*s in Group E_2 experienced the switch function first, those in Group E_3 experienced it in the third or middle position, and those in Group E_4

experienced it in the fifth or last position. The preutilization experiences were presented to *S*s as simple problems or tasks which they were to perform by using the switch. (No *S* encountered the relay prior to the two-string problem.) These *S*s were then confronted with the two-string problem. The five preutilization experiences encountered by the members of groups E_2, E_3, and E_4 follow: (*a*) Make a 2.5-volt bulb in a simple battery circuit go on and off (by connecting the microswitch into the circuit); (*b*) draw a design comprising five straight lines (by using the switch as a straightedge); (*c*) use the switch as a holder for a pipe cleaner (by inserting it in a small hole in the switch); (*d*) use the switch to hold a piece of cardboard in a tilted position (by using the switch as a support); and (*e*) use the switch as a paper-clip holder (by inserting it in the space provided by the switch contacts). The four nonspecialized functions were randomly ordered within a given series of preutilization experiences. It is assumed that the position of the switch function contributes more to the gestalt of the preutilization experiences than does the sequence of nonspecialized functions—just as the general shape and structure of a pencil sharpener contributes more to its meaning than does its color. That is, the essential function (switch) carries more weight in defining the gestalt than is true of the nonessential (nonspecialized) functions.

Buffer Task

To decrease any apparent connection between the preutilization experiences and the two-string problem, immediately after the preutilization experiences, each E-group *S* was asked to count from 1–50 as rapidly as possible. The experimenter (*E*) timed *S* and reported her "score" to her. *S* was then presented with the two-string problem. Members of the control group began with this task.

Two-String Problem

Two 3-foot strings were hung 8 feet apart from a light fixture 8½ feet above the floor. *S*'s task was to find a way to hold the two strings at the same time. The solution required that *S* tie an object (switch or relay) to one of the strings, set it swinging, grab the other string, and then catch the weighted string on the upswing. On a table in front of *S*, in a constant position, were the switch and the relay. All *S*s were able to make a choice of object and verbalize a solution within 6 minutes.

RESULTS

For purposes of analysis, the two-string problem was considered to be a one-item test. Each *S* who used the switch as a weight received a score of 1, and those who used the relay as a weight received a score of 0. Consequently, the lower the group mean score was, the more pronounced the functional-fixedness effect (i.e., the greater the tendency to choose the relay over the switch as a weight in the two-string problem).

It is recognized that a dichotomous variable does not allow one to meet certain assumptions underlying the analysis of variance model. However, the study by Boneau (1960) suggests that this may not have seriously distorted the analysis of variance results. Although chi-square is more appropriate, limited degrees of freedom do not allow the multiple comparisons desired in the present study (Lewis & Burke, 1949). Even so, the data were analyzed by both procedures with parallel results. The *one* statistical test most damaging to Flavell's hypothesis, the difference among E_2, E_3, E_4, and C, was significant ($\chi^2 = 9.91$, $df = 3$, $p < .02$).

A simple analysis of variance yielded an *F* ratio (3.12, $df = 4/110$) significant at less than the .02 level. Thus, there were reliable differences among the five groups. To specifically locate these differences, Duncan's new multiple range test was applied to the data (Edwards, 1960). The results of the multiple range test are presented in Table 2. The significant mean difference between the E_1 and C group is in agreement with previous studies and indicates that experiencing the switch in its usual function influenced the frequency with which it was used as a weight in the two-

TABLE 2

RESULTS OF THE DUNCAN MULTIPLE RANGE TEST AND NUMBER OF *S*s IN EACH GROUP CHOOSING THE SWITCH AND THE RELAY

No. *S*s choosing	Group				
	E$_2$	E$_1$	E$_4$	E$_3$	C
Switch	4	6	10	12	14
Relay	19	17	13	11	9
M	.17	.26	.43	.52	.61
	—	—	—		
		—	—	—	
			—	—	—

Note.—Any two *M*s underscored by the same line are not significantly different; any two *M*s not underscored by the same line are significantly different (p = .05).

string problem ($\chi^2 = 5.66$, $df = 1$, $p < .02$). This particular comparison was made to demonstrate that the method of assessing functional fixedness in the present study is comparable in results to those obtained by other investigators.

With respect to Flavell's hypothesis, groups E$_2$, E$_3$, E$_4$, and C should not have differed from each other. However, since Group E$_2$ differed from E$_3$ and C, it cannot be concluded that the *number* of unusual uses as such reduced the functional-fixedness effect in Flavell's study; as indicated by Group E$_2$, if *S* experienced the usual function *first*, functional fixedness was manifested even though four unusual uses were subsequently experienced.

DISCUSSION

As a positive phenomenon, functional fixedness reflects the ways in which the perceived properties or meanings of an object shift their emphasis depending upon previous experiences with and current demands of the object. The functional-fixedness experiment is a way of studying the ease or difficulty with which the properties of an object show themselves; or what is the same, it is a way of studying the object's psychological structure—the behavioral possibilities to which its aspects refer.

The results of the present study indicated that previous experiences with an object did not have an additive effect which operated irrespective of the total *structure* of these experiences. When *S* experiences the switch aspect of the switch, he has the experience that the object was probably manufactured for this purpose. This experience has a positive and a negative aspect. Positively, this experience renders the object less ambiguous—*S* discovers what it is. Negatively, this reduction of ambiguity narrows the background of possible meanings such that other functions become more remote and less discoverable. On the other hand, when *S* experiences the switch in a nonspecialized function (some minor or less-essential aspect), he realizes that this is not the function for which the object was manufactured, even though it realizes his purposes at hand. Hence, the background of discoverable meanings remains open. The meaning or perception of the object is not compellingly polarized in any one direction—the object remains ambiguous. *But* these experiences are modified by the total *structure* of experiences of which they are a part. Thus, when the essential aspect of the switch is emphasized first (E$_2$), the perceived properties of the object are polarized in the direction of this function (meaning), and subsequent experiences with some of its minor aspects do not disturb the emphasized properties—the discoverable meanings remain relatively hidden. When the minor aspects of the object are first emphasized (E$_3$, E$_4$), the background of discoverable meanings (the multiaspect nature of the object) is what is emphasized, and this emphasis is maintained even after the switch function is eventually experienced, and to about the same degree as is true when *S*s have not experienced the object in its specialized function (i.e., no differences among E_3, E_4, and C).

A consideration of the switch function apart from the total pattern of preproblem experiences would be misleading. Logically, the switch function can be said to provide the "same" experience to all *S*s irrespective of membership in a given experimental group; it is apparent, however, that the meaning or significance of the switch experience varied depending on the total structure of experiences to which it contributed as a part. This structure is precisely what varied from one experimental group to another.

Since the above descriptive interpretation is not in the style of contemporary American theoretical orientations, several alternative interpretations ought to be mentioned.

As Schulz (1960) has shown in a masterful fashion, functional fixedness can be viewed from the perspective of transfer. However, his position does not seem to account for the present results. Specifically, in relating the two-object functional-fixedness paradigm to his transfer model, Schulz (1960, p. 70) says:

> ...once the stimulus "relay" is associated with the response "electrical appliance" it is more difficult to attach a new response "weight" or "pendulum bob" to this stimulus.

Since, in the present study the switch was used in its electrical function in all experimental groups, the differential results obtained indicate that Schulz' position is at best incomplete. Although transfer theory does not seem to clarify the obtained results, it does have the advantage of specifying experimental variables to which functional fixedness can be related.

Kendler (1963) discussed functional fixedness from the perspective of habit competition. The habit to use the switch in its electrical function competes with the habit to use it as a weight. By using the switch as a switch, the "wrong" habit is strengthened; thus it is easier to use the relay as a weight in the subsequent problem situation. Kendler (1963, p. 366) further states that:

> ...when a habit to respond in a particular way to a given object has become very strong, it is difficult, if not impossible for us to respond to it in another way.

This position also fails to account for the differential results obtained. Further, no S had difficulty in *using* the switch in each of the four unusual ways even after experiencing the switch as a switch (E_2).

Another possible interpretation is provided by Judson and Cofer (1956). These authors present evidence for a primacy effect in reasoning which they interpret as indicating the availability of a habit family is modified depending on the first unambiguous stimulus encountered. If stimulus is broadened to include function or meaning in the sense of the present study, a crude parallel may be drawn between their study and the present one. Further, if the switch function is equated with Cofer's unambiguous stimulus, the parallel is a little closer. However, since all experimental groups experienced the switch, no differential degree of functional fixedness among groups would be expected. If, on the other hand, the experience of an unusual function first induces the habit family of "unusual uses," whereas the experience of the switch function first induces the habit family of "electrical uses," the data of the present study can be accounted for, but functional fixedness loses its meaning. That is, if *S* uses the switch in the two-string situation, it is because the unusual-uses habit has become fixated, while if *S* chooses the relay, it is because the electrical-uses habit has become fixated; both choices reflect fixedness of different types (usual and unusual) of functions. Even if this perspective were compelling, the fact that the primacy effect is itself in need of explanation reduces it to little more than a descriptive rather than explanatory position.

The present study called into question the validity of Flavell's hypothesis, and it emphasized the importance of the structure of previous experiences in terms of the significance such experiences may have in subsequent problem situations. If the phenomenological interpretation given seems to be directly in the tradition of gestalt psychology, it is because phenomenology has radicalized many of the fundamental ideas of this psychology (see, e.g., Gurwitsch, 1964; Merleau-Ponty, 1962, 1963).

REFERENCES

BIRCH, H. G., & RABINOWITZ, H. S. The negative effect of previous experience on productive thinking. *Journal of Experimental Psychology,* 1951, **41,** 121–125.

BONEAU, C. A. The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin,* 1960, **57,** 49–64.

EDWARDS, A. L. *Experimental design in psychological research.* (Rev. ed.) New York: Holt, Rinehart & Winston, 1960.

FLAVELL, J. H., COOPER, A., & LOISELLE, R. H. Effect of the number of preutilization functions on functional fixedness in problem solving. *Psychological Reports,* 1958, **4,** 343–350.

GURWITSCH, A. *The field of consciousness.* Pittsburgh: Duquesne Univer. Press, 1964.

JENSEN, J. On functional fixedness: Some critical remarks. *Scandinavian Journal of Psychology,* 1960, **1,** 157–162.

JUDSON, A. J., & COFER, C. N. Reasoning as an associative process: I. "Direction" in a simple verbal problem. *Psychological Reports,* 1956, **2,** 464–476.

KENDLER, H. H. *Basic psychology.* New York: Appleton-Century-Crofts, 1963.

LEWIS, D., & BURKE, C. J. The use and misuse of the chi-square test. *Psychological Bulletin,* 1949, **46,** 433–489.

MAIER, N. R. F. An aspect of human reasoning. *British Journal of Psychology,* 1933, **24,** 144–155.

MERLEAU-PONTY, M. *Phenomenology of perception.* (Trans. by C. Smith) London: Routledge & Kegan Paul, 1962.

MERLEAU-PONTY, M. *The structure of behavior.* (Trans. by A. L. Fisher) Boston: Beacon Press, 1963.

SCHULZ, R. W. Problem solving behavior and transfer. *Harvard Educational Review,* 1960, **30,** 61–77.

SCHUTZ, A. *Collected papers of.... The problem of social reality.* Vol. 1. The Hague: Martinus Nijhof, 1962.

VAN DE GEER, J. P. *A psychological study of problem solving.* Haarlem, Netherlands: de Toorts, 1957.

(Received August 30, 1965)

What if they aren't "average"

The NEW ROCHESTER OCCUPATIONAL READING SERIES from SRA has been invented for young people whose reading abilities are between the second and fifth grades.

"When a youngster reaches his teens with only a second or third grade reading ability, as so many of our retarded and deprived children do, 'John and Jane' readers are a sneer and a delusion. In the early days of this century when we taught reading to adult immigrants we had to invent new materials geared to adult state. Now for our under-equipped teen-agers we have to invent reading materials that are geared to their interests and (let's face it) that means the getting of jobs to get money to get dates and/or jalopies and sports equipment, all of which have to do with the urge toward adult status."*

Stories in the SERIES are built around situations of genuine interest and concern to older students; namely, employment possibilities and opportunities, and the use of leisure time. The NEW ROCHESTER OCCUPATIONAL READING SERIES is not sophisticated in nature, nor is it expected to fill all the educational gaps in the life of the disadvantaged or the slow learner. Rather, this reading series exists to help improve their language and reading skills, and to help them become employable.

*James M. Spinning in "Nation's Schools," April 1964. Copyright 1964 by McGraw-Hill, Inc. All rights reserved.

Send me your new free folder containing program information and prices of the NEW ROCHESTER OCCUPATIONAL READING SERIES.

NAME____________________

POSITION____________________

SCHOOL____________________

ADDRESS____________________

CITY____________________

STATE________________ZIP______

S R A

Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A subsidiary of IBM

A question for many cultures . . .

How does a child learn?

SOCIAL FOUNDATIONS of EDUCATION

A Cross-Cultural Approach

By COLE S. BREMBECK, *Michigan State University.* Invaluable to any prospective teacher, this book describes in depth how the social character of the school determines how students learn. The author treats the social environment within a school as an expression of the larger society without, and shows how a teacher can take cues from his students' environment about what to teach and how.

Avoiding useless generalizations, the text abounds with stories, illustrations, and specific examples put together in a lively, readable style that will help to identify a student's cultural values and aspirations. The author provides four detailed case studies that serve to compare widely divergent environments. He analyzes student cultures in a city slum, a suburban community, a midwestern town, and a Pakistan village—drawing useful conclusions from each. *1966. Approx. 544 pages. Prob. $7.50.*

And now in paperback . . .

The SEARCH for ABILITY

Standardized Testing in Social Perspective

210 pages. *$1.45.*

JOHN DEWEY as EDUCATOR

His Design for Work in Education (1894–1904)

By ARTHUR G. WIRTH, *Washington University, St. Louis.* A study, based on primary sources, of the educational theory and practice John Dewey represented during the one period of his career when he was a practicing educator. The author examines the evolution of Dewey's philosophical and psychological theory against the practice developed in the Laboratory School founded by Dewey at the University of Chicago. *1966. Approx. 324 pages. Prob. $6.50 (cloth) and $3.95 (paper).*

ROLE THEORY

Concepts and Research

Edited by BRUCE J. BIDDLE, *University of Missouri; and* EDWIN J. THOMAS, *The University of Michigan; with a foreword by* Theodore M. Newcomb. The first book devoted exclusively to the emerging specialization of role theory. There are four original essays which cover the history and current status of role theory as well as the framework of terms and concepts in the field; 47 reprinted selections cover significant theory and specialized research topics. *1966. Approx. 416 pages. Prob. $8.95.*

JOHN WILEY & SONS, Inc.

. . . Wiley has many answers.

COGNITION and THOUGHT

An Information Processing Approach

By WALTER R. REITMAN, *The University of Michigan.* An introduction to computer simulation and an analysis of information processing concepts and their psychological implications. Requiring no prior knowledge of computers, the book features a discussion of selected problems in cognition and thought, carefully organized to provide a coherent picture of information processing theories. *1965. 312 pages. $7.95.*

DYNAMICS OF RESPONSE

By JOSEPH M. NOTTERMAN, *Princeton University; and* DONALD E. MINTZ, *City College of the City University of New York.* Extends the analysis of instrumental conditioning to the measurable properties of individual responses. Peak force of response is given particular attention, and all responses are studied as they are systematically influenced by both conventional and novel experimental procedures. *1965. 277 pages. $8.95.*

CULTURE in AMERICAN EDUCATION

By RUTH LANDES, *McMaster University.* Describes how a California teacher-training experiment advanced teaching and social work by acquainting public school educators with cultural traits of the state's minority groups (especially Negro and Mexican-American). The author discusses Claremont College's unique project designed to familiarize educators with the social complexities of dominant and minority groups, so that teachers could evolve goals and techniques accordingly.

The training stressed comparative studies of the family, schooling, health care, and other institutionalized concerns in diverse sociological settings. The results proved that the most direct way to tap students' energies is through an understanding of their group customs. *1965. 330 pages. $7.95.*

GROUP VALUES THROUGH CHILDREN'S DRAWINGS

By WAYNE DENNIS, *Brooklyn College of the City University of New York.* Presents a unique approach to the determination of cultural values in a given society through the interpretation of children's drawings. The author examines the drawings of children from various socio-economic groups in thirteen countries. He supplements his research and conclusions with several of the drawings obtained in the study. *1966. 211 pages. Illus. $6.95.*

DESCRIPTIVE STATISTICS

A Programmed Textbook: In Two Volumes

By LASSAR G. GOTKIN *and* LEO S. GOLDSTEIN, *both of the Institute for Developmental Studies, New York Medical College; formerly of the Center for Programmed Instruction.* Each unit of this two-volume text is organized so that the reader can proceed from basic knowledge to complex applications. This is done by actively involving him in the learning processes.

Volume 1: *1964. 255 pages. $3.95.*
Volume 2: *1965. 229 pages. $3.95.*

605 Third Avenue, New York, New York 10016

BLAISDELL
BOOKS IN PSYCHOLOGY

Consulting Editor: Raymond G. Kuhlen,
Syracuse University

In preparation . . .

READINGS IN THE PSYCHOLOGY OF CHILDHOOD AND ADOLESCENCE
by William J. Meyer, Syracuse University

EDUCATIONAL PSYCHOLOGICAL MEASUREMENT:
Contributions to Theory and Practice
by David A. Payne, Syracuse University and Robert F. McMorris, SUNY at Albany

TOPICS IN EDUCATIONAL PSYCHOLOGY

A new series of original works in paperbound editions.

Forthcoming . . .

THE PSYCHOLOGY OF VOCATIONAL CHOICE
by John L. Holland, The American College Testing Program, Iowa City

A PRIMER IN STATISTICAL PRINCIPLES
by Victor E. McGee, Dartmouth College

In preparation . . .

COGNITIVE DEVELOPMENT
by Charles D. Smock, Purdue University

A PHENOMENOLOGICAL INTERPRETATION OF THE EDUCATIVE PROCESS
by Donald Snygg, State University of New York, College at Oswego

PSYCHOLOGY OF ADOLESCENCE
by W. Cody Wilson, Advanced Research Projects Agency, Washington, D.C.

BLAISDELL PUBLISHING COMPANY
A Division of Ginn and Company
275 Wyman Street, Waltham, Mass. 02154

from
RANDOM HOUSE

In preparation . . .
THREE BABIES
Biographies of Cognitive Development
Edited by JOSEPH CHURCH, *Brooklyn College of The City University of New York*
May 1966; about 288 pages; about $2.75 paperbound; about $4.25 hardcover
RANDOM HOUSE STUDIES IN EDUCATION
A new series of scholarly works in paperbound covers.
In preparation . . .
GUIDING THE GIFTED CHILD
by GEORGE THOMAS, *State Department of Education, Albany, New York and* JOSEPH CRESCIMBENI, *Jacksonville University*
Summer 1966; 160 pages; about $1.95
Recently published
THEORIES OF ADOLESCENCE
by ROLF E. MUUSS, *Goucher College* PP 22; $1.95

For further information, write to

The College Department
RANDOM HOUSE, INC.
501 Madison Avenue
New York 10022

Publication Manual

of the American Psychological Association

1957 Revision

A revision of the 1952 Manual, detailed instructions are given for the preparation of scientific articles. Organization and presentation of tabular material, figures and graphs, and reference lists are included. Scientists writing for publication will find the Publication Manual an indispensable guide.

Price $1.00

Order from
AMERICAN PSYCHOLOGICAL ASSOCIATION
1200 Seventeenth Street, N.W., Washington, D. C. 20036

Reading in High Gear

Reading in High Gear is a new approach to the problem of teaching reading; this program assumes only that the student can speak English.

Reading in High Gear has been developed by Dr. Myron Woolman, primarily for the culturally disadvantaged adolescent and adult. Dr. Woolman's program is based on his Progressive Choice Learning Theory, first validated in his study "The Effect of Varying the Number of Choices in the Identification of Very Similar Stimuli."*

The two basic principles of Progressive Choice are: Discrimination . . . two things taught in sequence should be dissimilar in appearance to minimize confusion and facilitate learning; Invariance . . . in the initial learning steps, only one response will be required for any one element taught.

Reading in High Gear is particularly appropriate for teachers looking for suitable material to meet individual abilities of remedial and pre-dropout students. For further information about Progressive Choice and **Reading in High Gear,** see the noted articles and return the coupon to SRA.

*1. Woolman, Myron, "The Effect of Varying the Number of Choices in the Identification of Very Similar Stimuli," unpublished doctoral dissertation. Columbia University, 1955. Complete microfilm copy available from University Microfilms, Ann Arbor, Michigan.

2. "New Teaching Technique Raises Reading Level," **Jet,** August 26, 1965.

3. Bloomer, R. H., "Reading Methodology: An Experimental Test of Some Alternative Organizational Principles," **The Reading Teacher,** January 1961.

4. Rothberg, I. C. and Woolman, M., "Verbal Paired-Associate Learning as a Function of Grouping Similar Stimuli or Responses," **Journal of Experimental Psychology,** January 1963. Vol. 65, No. 1.

Please send a free copy of Dr. Woolman's twenty-page booklet about **Reading in High Gear.**

Name____________________

Position____________________

School____________________

Address____________________

City__________ State__________ Zip__________

S R A

Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A subsidiary of IBM

Challenging . . .

CYBERNETIC PRINCIPLES OF LEARNING AND EDUCATIONAL DESIGN

KARL U. SMITH, University of Wisconsin

MARGARET FOLTZ SMITH

An outstanding text that relates human learning and educational skills to basic cybernetic mechanisms of behavior

Challenging many of the tenets of conventional learning theory, *Cybernetic Principles of Learning and Educational Design* constitutes a major breakthrough that provides a new basis for studying human learning that will be significant in the classroom, as well as in the laboratory. The major premise set forth in the book is that the learning situation *must be suited to the control capabilities of the learner.*

To illustrate the implications of cybernetic theory for learning, the authors . . .

- describe the evolutionary development of human learning
- illustrate the ways in which man has become adapted to his own educational design
- emphasize research in meaningful educational settings
- survey methods of teaching and training
- summarize the results of investigations into their effectiveness

January 1966 **542 pp.** **$8.95**

Holt, Rinehart and Winston, Inc.

383 Madison Avenue, New York, N.Y. 10017

New books of interest

A PRACTICAL INTRODUCTION TO MEASUREMENT AND EVALUATION

Second Edition

H. H. REMMERS, N. L. GAGE & J. FRANCIS RUMMEL Retaining the book's original organization, the Second Edition focuses on the theories and techniques of statistical measurement; the administration, scoring, and interpretation of tests; the evaluation of classroom instruction through teacher-made tests; and the appraisal of emotional and social adjustment. There is a new section, "Taxonomy of Educational Objectives: Affective Domain," and discussions of the status of the measurement of creativity and of *Digitek*, the new scoring machine. 390 pp. $6.95

SECONDARY REINFORCEMENT

Selected Experiments

EDWARD L. WIKE This is the first full integration of present knowledge regarding positive secondary reinforcement. The text isolates functional relationships and exemplifies them by presenting an organized set of experimental studies in their complete, original form. The Introduction and Resume consider such topics as Hull's Approach to Motivation, Skinner's Approach to Motivation, Paradigms for Secondary Reinforcement, and Principles of Secondary Reinforcement. 516 pp. $13.75

Harper & Row Publishers

49 East 33d St., New York 10016

INTERVIEWERS
MARKET RESEARCH
PRIMARILY IN THE MEDICAL FIELD

* * *

Part-time work, assignments adjustable to available time: Ph.D. preferred, M.A. considered: Interviewing experience required, but medical experience not essential: Any location in the United States. High per interview rate plus bonus.

* * *

Address inquiries to:

THE MENDOTA
RESEARCH GROUP
31 SYLVAN AVENUE
ENGLEWOOD CLIFFS,
NEW JERSEY

New and recent Prentice-Hall books

AN INTRODUCTION TO PERSONALITY:
A Research Approach

by DONN BYRNE, University of Texas. Concentrating on current aspects of research which deal with personality as a science rather than as a branch of speculative philosophy, this new basic text provides an understanding of the way in which the scientific method is utilized in studying human behavior. *April 1966, 560 pp., $7.75*

PERSONALITY RESEARCH: A Book of Readings

edited by DONN BYRNE, University of Texas, and MARSHALL L. HAMILTON, San Jose State College. Here is a representative picture of current research, theoretical formulations, and issues typical of the field of personality. Ideal when used in conjunction with the above title. *April 1966, 432 pp., paperbound, $5.50*

THE SELF IN GROWTH, TEACHING AND LEARNING:
Selected Readings

edited by DON E. HAMACHEK, Michigan State University. An anthology concentrating specifically on the self as it is influenced by growth, teaching, learning and perception. Some of the articles reflect both the theoretical and philosophical undercurrents beneath the self as a psychological frame of reference. *1965, 576 pp., paperbound, $7.50*

MEASURING EDUCATIONAL ACHIEVEMENT

by ROBERT L. EBEL, Michigan State University. A practical approach to the problems and procedures involved in constructing and using classroom tests of educational achievement. The book offers very concrete suggestions for planning, constructing, administering, scoring, and analyzing the results of these tests, and features an extensive discussion of the functions of such tests. *1965, 481 pp., $7.95*

PSYCHOPATHOLOGY OF CHILDHOOD

by JANE W. KESSLER, Director, Mental Development Center, Western Reserve University. This new book discusses child behavior problems at all degrees of seriousness, and combines research findings and clinical observations for each. It is *problem-centered* rather than *profession-centered. February 1966, 544 pp., $8.25*

TEACHING DISADVANTAGED CHILDREN IN THE PRESCHOOL

by CARL BEREITER and SIEGFRIED ENGELMANN, both at the Institute for Research on Exceptional Children, University of Illinois. June 1966, approx. 352 pp., $7.95

(PRICES SHOWN ARE FOR STUDENT USE)

For approval copies, write: Box 903

PRENTICE-HALL, Englewood Cliffs, N.J. 07632

BORZOI College Textbooks

EDUCATIONAL PSYCHOLOGY

An Application of Social and Behavioral Theory

by LOUIS M. SMITH *and* BRYCE B. HUDGINS

both of Washington University at St. Louis

> One of the finest books in this area I have ever seen. Its comprehensiveness is delightful! . . .
>
> —Harry V. Barnard, *Rutgers University*

1964; 576 pages; 21 figures; 73 tables; $8.50

An **INSTRUCTOR'S MANUAL,** prepared by Louis M. Smith and Bryce B. Hudgins, is available to accompany EDUCATIONAL PSYCHOLOGY.

For further information, write to:

ALFRED A. KNOPF, *Publisher*

College Department **501 Madison Avenue** **New York 10022**

The 1965 DIRECTORY of the AMERICAN PSYCHOLOGICAL ASSOCIATION

A revised standard reference containing biographical and geographical listings of the Association's 23,500 fellows, members, and associates.

A valuable tool for psychologists and members of related professions
libraries publishers
personnel offices

$10.00 ($5.00 to APA members)

Order Form

American
Psychological
Association
Order Department
1200 17th Street, N.W.
Washington, D.C. 20036

Date. . . . .

Please send. . . . . . .copies of the 1965 APA Directory to:

. . . .
(please print or type)

. . . .

. . . .

Check enclosed. . . .
Please bill. . . .

Signature . . . .

announcing the Second Edition

LEARNING AND HUMAN ABILITIES

EDUCATIONAL PSYCHOLOGY

Herbert J. Klausmeier · William Goodwin

Almost completely rewritten and thoroughly updated in the Second Edition, this widely adopted text continues to emphasize the concept of emerging human abilities —thus integrating the notions of growth and learning that are traditionally treated separately. The new content and organization meet a wider range of needs of students and professors. Especially suited to courses which emphasize teaching-learning processes in school settings as well as the development of intellectual abilities the book is ideal for several types of courses: a teaching-oriented course, a science-oriented course, or a comprehensive course which includes measurement and experimental design. Course outlines have been prepared for the instructor, and suggest a variety of combinations of all or parts of each of the 18 chapters. 720 pp. $8.95. *Just published.*

A Student Workbook and Adjunct Program, by Herbert J. Klausmeier William L. Goodwin, and Robert Conry is available for use with the text. This combined supplement consists of a student workbook, comprised of questions that encourage high-level intellectual abilities of the students (including reasoning, evaluation, and creativity), and an adjunct program, providing forty multiple-choice items and subjective questions. **A Student Evaluation Guide**, providing a test-item file, is available to teachers.

READINGS IN LEARNING AND HUMAN ABILITIES

EDUCATIONAL PSYCHOLOGY

Richard E. Ripple, Editor

This volume contains forty-nine unabridged readings for use with LEARNING AND HUMAN ABILITIES: *Educational Psychology* or with other introductory educational psychology texts. A cross-reference chart correlates the readings with chapters in sixteen current texts. Although emphasis is on articles reporting empirical research studies, some theoretical articles and survey articles reviewing research are included. A balance is achieved among articles dealing with the elementary, secondary, and college levels. 586 pp. $4.95/paper

HARPER & ROW, PUBLISHERS • 49 E. 33d ST., NEW YORK 10016

FOUND!

825 copies of

SCHOOL PSYCHOLOGISTS AT MID-CENTURY

Norma E. Cutts, Editor

The printer has discovered on his storage shelves over 800 copies of this valuable report—supposedly out of print since 1961. To meet the continuing demand for this popular book, it is offered again—while the copies last.

This Report on the Thayer Conference is an intensive study on the functions, qualifications, and training of school psychologists. It was originally published in 1955 and reprinted in 1958.

230 pages **Price: $2.75**

ORDER FROM:

American Psychological Association
1200 17th Street, N.W.
Washington, D.C. 20036

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 57 June 1966 Number 3

CONTENTS

Intellectual Status and Intellectual Growth 121

ROBERT L. THORNDIKE

Conditions of Prequestioning and Retention of Meaningful Material 128

D. E. BERLYNE

Reversals in Children as a Function of Midline Body Orientation 133

ROBERT W. ZASLOW

Effects of Delay of Information Feedback and Length of Postfeedback Interval on Linear Programed Learning 140

FREDERICK J. BOERSMA

Eight Professors—A Study on College Students' Preferences among their Teachers 146

KAORU YAMAMOTO AND HENRY F. DIZNEY

Effects of Written Verbalization and Timing of Information on Problem Solving in Programed Learning 151

ROBERT J. SEIDEL AND IRIS C. ROTBERG

Effect of Experimentally Learned Word Associations on the Acquisition of Reading Responses 159

S. JAY SAMUELS

Concept Development in Piagetian Terms in Relation to School Attainment ... 164

P. S. FREYBERG

Generalized Effects of Praise and Reproof 169

HARRY E. ANDERSON, JR., WILLIAM F. WHITE, AND JAMES A. WASH

Emotionally Loaded Words and the Acquisition of a Sight Vocabulary 174

DAVID R. OLSON AND A. S. PAU

Journal o Educational Psychology
1966, Vol. 57, No. 3, 121–127

INTELLECTUAL STATUS AND INTELLECTUAL GROWTH[1]

ROBERT L. THORNDIKE
Teachers College, Columbia University

Studying growth in ability is made difficult by (a) unreliability of measures, (b) inequality of units in which measures are expressed, and (c) nonequivalence in what is measured by different instruments and at different levels. These problems are illustrated in analyses of Harvard Growth Study data, where 2 intelligence tests were given to each of 593 pupils each yr. for 6 yrs. The maximum reliability of intellectual growth appears to be about .25 or .30. An unbiased estimate of the obtained correlation of status with gain appears about .10, but this correlation has apparently been reduced by inequalities in the units of measure between (and within) instruments, and by the fact that different tests measure somewhat different functions.

Ever since psychologists began investigating individual differences, they have been interested in determining the extent to which differences in intellectual growth patterns are orderly and predictable from something that can be known about the individual relatively early in life, to what extent these differences are under the control of ascertainable and describable environmental conditions, and to what extent they are inherently erratic and unpredictable.

Much of the use of tests in education is based on the premise of stability—that the bright child at Age 8 will continue to be a bright child at Age 10 or 12. And there is abundant evidence that, within limits, this is the case. But this is stability of status—that he *is* still bright, not necessarily that he has continued to grow rapidly. As Anderson pointed out as early as 1939, a good deal of stability of intellectual *status* can be accounted for by a model that says that ability at a later date consists of ability at an earlier date plus a growth increment that is completely uncorrelated with the earlier level of ability. Bloom (1964) has developed this theme at some length in his recent book, *Stability and Change in Human Characteristics.*

What *are* the substantive findings with respect to increments in intellectual ability? Do these increments represent an extrapolation of growth trends established at an early age? Or do they represent the addition of a random increment—random in the sense that it is completely uncorrelated with intellectual status at the beginning of the period in question?

Unfortunately, it is very hard to say. The study of change is one of the trickiest areas in which to work if one must use the fallible instruments that are the psychometrician's tools. The questions asked are simple ones: To what extent are the children who show rapid intellectual growth over one period of time the same children who had a prior history of rapid intellectual growth? To what extent are individual differences in rate of intellectual growth over an earlier period of time related to individual differences in growth rate over some later period? To simple questions we would like simple answers, and that might be possible if three conditions could be met: having completely error-free measures, having measures expressed in meaningfully equal units, and having at all ages measures referring to identically the same attribute of the individual. Then gains could be correlated with initial status, and it would be possible to see to what extent initial status is related to gain or gain over one period is related to gain over another period. In the best of physical measures—that is, measures of height or weight—these conditions are approached, but when we deal with mental tests, we are distressingly far away from meeting the specifications.

---

[1] This paper was presented in a slightly different version as a presidential address before the American Psychological Association's Division of Educational Psychology (Division 15), Chicago, 1965.

Three problem areas will be examined in turn—reliability of measurement, equality of units, and comparability of function—to see how serious they turn out to be, and what the implications are for inquiry.

RELIABILITY OF MEASUREMENT

Much of this presentation will be based on some reanalyses of Harvard Growth Study data. In terms of extensiveness and continuity of testing these data are in many respects unique, but they were gathered when computing facilities were neolithic by present-day standards, and it is the author's impression that they have never been fully analyzed.

In the Harvard Growth Study that Dearborn initiated in the early 1920's, over 3,000 pupils in public schools near Boston were studied from the time they entered the first grade until they left the school system—either by dropping out, moving away, or completing the twelfth grade (Dearborn, Rothney, & Shuttleworth, 1938). A great many different anatomical and physical measures were obtained, but interest here is limited to the group-intelligence-test results, which were obtained by first one and then two different tests each year throughout the duration of the experiment.

The present analysis is limited to the 593 cases out of the total of some 3,000 for whom complete records were available on one intelligence test at Age 8, and on two tests at each age from 9 through 14. These are, of course, group tests—the tests that were available for use in the 1920's—with whatever limitations attach to them. In all, 13 test scores expressed in mental-age units were available, spread over a 6-year period. The matrix of correlations was computed not only among the 13 status scores but also among all the difference scores generated by taking the tests in pairs. Thus the correlation of Test 9_1 with Test 10_1 and the correlation of 9_2 with 10_2 is available, together with the correlation of the difference between 9_1 and 10_1—the gain as measured by this one pair of tests—with the difference between 9_2 and 10_2—the gain as measured by another pair of tests.

With these data in hand, the problems arising from errors of measurement will first be examined. Psychologists are all quite conscious of the fact that intelligence tests fall well short of being perfectly reliable. When it is thought of at all, a correlation coefficient of .85 or .90, or, most optimistically, .95, is conceived as representing the reliability for a group homogeneous with respect to age or grade. And psychologists are all vaguely aware that change scores are less reliable than status scores. It is recognized that much of what is measured at any one point in time is common to what is measured at another point in time, and that this common component disappears when gain scores rather than status scores are used. But how *much* less reliable are the gain scores? What *is* the residual reliability of a set of change scores?

To answer this question empirically, there is within the Harvard Growth Study data a whole set of correlations between *two* gain scores, each covering the same span in the child's life, but each being based on a separate pair of tests. That is, gain based on Tests 9_1 and 10_1 can actually be correlated with gain based on 9_2 and 10_2. Some of these gains cover only a span of 1 year—9–10, 10–11, 11–12. Some cover 2 years, some 3, some 4, and one pair of correlations is for gains over 5 years. The actual average correlations are as follows: 1-year interval, .101; 2-year interval, .240; 3-year interval, .266; 4-year interval, .188; 5-year interval, .265.

It appears, then, that the ceiling for the reliability of a measure of intellectual growth based upon the administration of two group-intelligence tests, even with an interval of several years between them, is somewhere around .25 or .30. The result for individual tests might be somewhat higher, but probably not much, since the data being examined involve children 9 or 10 years old at the earlier testing. In any study involving gain scores, it must be recognized that the datum is really fragile, with a reliability that appears to lie in the range from .1–.3, depending upon the tests and the interval, and research must be planned and evaluated accordingly.

As has been repeatedly pointed out by previous critics of growth studies, the substantial amount of error variance in test scores leads to especially misleading conclusions if one attempts to correlate gains with initial status, using the same testing to provide both the initial-status score and the subtrahend in the gain score. The reason, of course, is that the error of measurement in this common test appears in the status score with a positive sign and in the gain score with a negative sign, and so generates a spurious tendency towards negative correlation based on the reversed error component. That is, if status is represented by Test A, and gain by B minus A, a positive measurement error in A results in a falsely high A score, but a falsely low gain.

The present data provide an excellent opportunity to demonstrate this effect empirically and to get some idea of its size. It is possible to do this by comparing two sets of correlations between initial status and subsequent gain. In one set, the same testing is used both to define status and to calculate gain; in the other set gains are calculated from the second set of measures. That is, if status is defined by Test 9_1, in the first case gain is defined as 10_1 minus 9_1, while in the second case gain is defined as 10_2 minus 9_2. These two types of correlations are shown in Table 1.

The average value varies somewhat with the time interval, but a roughly representative figure might be −.20 when the same test is used to define both status and gain and .10 when different tests are used in the two cases. Thus, the difference between −.20 and .10 is the magnitude of the spurious effect generated by lack of experimental independence in the data and the incorporation of a common measurement error with opposite sign in each variable.

TABLE 1

CORRELATION OF INITIAL STATUS WITH GAIN

Based on common test		Based on experimentally independent tests	
Intervals (in yrs.)		Intervals (in yrs.)	
1	−.273	1	.045
2	−.287	2	.006
3	−.283	3	.031
4	−.194	4	.139
5	−.053	5	.329

The experimentally independent correlations just cited, for which the typical value is perhaps .10, provide the best estimate of the correct and unbiased correlation between initial status and subsequent gain. It is a pretty minute correlation, but considering what has just been seen about the reliability of gain measures, its small size is not too surprising. A rough estimate can be made of what it would become if it were corrected for the unreliability of both status measure and gain measure. Using .85 and .25 as working estimates of the two reliabilities, the corrected correlation becomes about .22—still pretty unimpressive.

In the light of the results to this point, two conclusions seem reasonable. The first is that rate of gain in intelligence-test score over the school years is in large part inherently unpredictable because of the extremely low reliability of the measures of individual differences in gain. The second conclusion, which becomes rather tentative because of the situation implied by the first conclusion, is that the nonchance component in gain score is positively related, but only slightly positively related to initial status. In considerable part, the factors that produce gains during a specified time span appear to be different from those that produced the level of competence exhibited at the beginning of the period.

EQUALITY OF UNITS OF MEASURE

But it will be remembered that three requirements were stated for meaningful studies of individual differences in mental growth. So far only the effects of violation of the first condition have been considered—error-free measures. What about equality of units and homogeneity of the function being measured?

Whenever correlational techniques are applied to psychological data, it is assumed that the numerals in which the variables are expressed represent equal increments in some attribute. It is also recognized that this assumption is usually not well supported. But for "rough and ready" studies of relationship, the violation of the assumption usually does not hurt much. However, when starting to deal with something as fragile as a change score, the violation of this basic assumption becomes a good deal more critical. The variance of change scores is in some cases no more than one-fourth the size of the variance of status scores. With two measures contributing to the gain score, any variance arising from irregularity in the two tests' converted score will tend to bulk 8 times as large as a fraction of the gain-score variance.

Furthermore, any systematic shift in the size of the units, such as that produced by a test ceiling too low to permit full expression of the abilities of those tested, may produce a systematic distortion of the relationship between status and gain. With such a ceiling effect those initially highest will not be fully measured by the test at the later ages, and will necessarily show smaller increments. Thus, in the Harvard Growth Study all test results were presented in mental-age units. However, for the older groups of pupils—13- and 14-year-olds—the mental-age equivalents for the brighter fraction were necessarily extrapolated values, based on whatever type of assumption may have seemed reasonable to the test makers of the 1920's. (The test manuals are notably terse on the procedures used to set up the scale of mental-age equivalents.) If the extrapolations were made somewhat conservative, so as not to overstate a level of ability that was arbitrary and not empirically based on the age schema then in favor, the effect would be to underestimate the scale values for able pupils at the higher ages, with the result that gain scores for these pupils would be depressed. Even a small systematic depression could substantially reduce or even eliminate the necessarily small correlation between initial status and the highly unreliable gain score.

Some evidence from the present data bearing on the reality of the problem of unit of measure comes from the comparison of means and standard deviations (expressed as mental ages) of the two tests given in the same year. These are shown in Table 2. Differences in parameters for these nearly simultaneous tests cannot be attributed to growth or environmental impact or anything but the characteristics of the tests themselves. Thus, it is found that the two 9-year test means of 9-10 and 10-0, respectively, agree quite well; but for the corresponding standard deviations of 20 and 13 months, one surpasses the other by 50%. For the 10-year tests the standard deviations agree well—both are 23 months—but the two means are 11-5 and 10-3, a difference of 14 months. At 11 there is a 6-month difference between the two means and a 5-month difference in the two standard deviations. Similar, though less dramatic, differences are found at 12, 13, and 14. Clearly, the units of the *different* tests are not equivalent. One suspects that the same can be said for different segments of the score scale for a single test.

TABLE 2

MEANS AND STANDARD DEVIATIONS OF PAIRS OF TESTS GIVEN IN THE SAME YEAR

Ss	Test No. 1		Test No. 2	
	M	*SD*	*M*	*SD*
9-year-olds	118.0	19.9	119.7	13.3
10-year-olds	137.4	23.0	123.2	22.7
11-year-olds	153.6	24.4	148.1	29.4
12-year-olds	164.6	22.9	164.5	26.3
13-year-olds	167.7	22.0	176.1	24.7
14-year-olds	178.1	23.9	181.4	23.0

The changing standard deviation for different variables makes for quite erratic correlations between status and gain. Simple algebra reveals the fact that the sign and the size of the correlation of Status Score A with gain from Score B to Score C depends upon the relative size of $s_C r_{AC}$ and $s_B r_{AB}$. The formula is as follows:

$$r_{A(C-B)} = \frac{r_{AC}s_C - r_{AB}s_B}{\sqrt{s_C^2 + s_B^2 - 2r_{BC}s_Bs_C}}.$$

Thus a change in scale unit, resulting in a change in standard deviation, can wreak havoc upon one's results. This appeared, in our data, in the wide range of values that were obtained for the correlation of status with gain. Though a value of .10 has been suggested as best representing the relationship, this average value is based on a range of correlations from .64 to −.43 for different specific combinations of initial test and subsequent test to define gain. This variation illustrates the point that the result obtained is spectacularly responsive to distortions of the unit of measure as between the initial and final test.

One particularly misleading pastime that is sometimes engaged in, even by those who should know better, is to study the relationship between an initial IQ and the change in IQ during some subsequent period. The absurdities of this procedure become clearer if the IQ is thought of as a standard score keeping the same mean and standard deviation at each age—a characteristic that is explicitly true of recently developed tests and approximately true of those developed in an earlier day.

When the standard deviations of two sets of scores at different points in time, B and C, are forced to be equal, and the correlation of a third score, A, with the difference between the two is studied, the sign of the resulting correlation depends solely on the relative size of the two correlations r_{AB} and r_{AC}. Thus, if it is desired to study the correlation between IQ on Test A at Age 8 and change in IQ from 9–12, the only way that the correlation *can* be positive is for the 8-year-old test to correlate more highly with the 12-year-old test than with the 9-year-old test—a fairly improbable and unnatural situation. If the correlation with the later test is lower, correlation with the change score will necessarily be negative.

By eliminating from the score scale the differences in standard deviation at different ages, that which is the essence of growth is eliminated—the greater variability of specimens as they mature. Imagine a group of adults whose heights and weights showed no greater standard deviations than those of newborn babies! And a statistical treatment that automatically excludes greater variability in intellect as we go from birth to maturity is equally absurd. Thus, any studies of intellectual growth that use as a score scale an IQ with a uniform standard deviation from age to age are, by that very fact, meaningless for any analytical study of the growth process. The constraint that has been put on the score scale assures distorted results.

UNIFORMITY OF FUNCTION MEASURED

A basic assumption, if growth is to be studied, is that the same function is being measured throughout the range of ages being studied. But when intellectual performance over a range of ages is studied, it is quite usual for the particular test to change as we go from an earlier to a later age. The change in test is likely to be accompanied by some change in the functions measured by the test, depressing the consistency between early and late performance, with a resulting depression of the relationship between a status measure based on Test A and a gain measure based on Tests B and C.

The Growth Study data permit a look at this problem because there were several age spans during which the same test was used in successive years. In all, six comparisons were possible between the correlation of some test, A, used in one year with Test A in the following year and its correlation with a different test, B, the following year. In the same way, six comparisons were possible of correlations of some test, A, in a given year with Tests A and B in the preceding year. The same-test correlations with a 1-year interval averaged to exactly .80, the correlations with another test in the following year to .75, and with another test in the prior year, .73, or a pooled average of .74 for the correlation between different tests.

This means that, while identical tests showed 64% of common variance from one year to the next, with a change in test the common variance was only 55%, or only 86% as great. Nine percent of the total variance and 14% of the stably measured variance was unique to the specific test.

If this change in function measured was cumulative, so that each time a new test was introduced 14% of what was measured by the prior test was replaced with new variance unique to the new test, it would take only seven shifts before measures of completely unrelated functions were obtained. Of course, this is not the way things operate. As we go from Test A to B to C, it is likely that part of what is unique to Test C so far as Test B is concerned is shared in common with Test A. It is unlikely that a clear picture of further drop in correlation with a successive shift will be found. Thus, in the Harvard data, Dearborn C at Age 11 correlates .808 with Dearborn C at Age 12 and only .760 with Haggerty Delta 2 at Age 12; .693 with Haggerty Delta 2 at Age 13, but .685 with Terman Group at Age 13, and .736 with Terman Group at Age 14. There is no evidence of further drop as we move on to the third test, in this case the Terman Group Test.

The reduction in correlation from .80 to .74 that has just been seen in comparing same with different tests does not seem like very much, but once again reference to the formula for correlation between one status score and another change measure shows that anything that introduces even a small extraneous reduction in the size of the correlation between status and the terminal measure of the change sequence can extinguish or even reverse the sign of the correlation between status and change. Thus, though the lower correlation between different tests looks pretty unimportant as we look at the reasonably robust test scores, it becomes very significant when we are dealing with the very fragile change scores. A drop from .80 to .74 in intercorrelation could reduce the correlation between status and gain from .20 to .10, cutting it in half, or from .10 to 0, existinguishing it completely.

CONCLUSION

Thus it has been seen that change scores are very sensitive creatures, very responsive to changes in score scale and to changes in the function being measured. So perhaps we need to modify the earlier conclusion that seemed appropriate from the Harvard Growth Study data—the conclusion that only a small fraction of the gain in later competence can be thought of as arising from a continuation of early growth trends. Perhaps it must be concluded that, unless and until scales are developed that are truly homogeneous in the functions that they tap at all levels and are truly expressed in equal units, we will have to forego serious attempts to give a quantitative answer to the simple but tantalizing question: To what extent will the children who have grown rapidly in intellect up to the present moment continue to grow rapidly in the future?

REFERENCES

ANDERSON, J. E. The limitation of infant and preschool tests in the measurement of intelligence. *Journal of Psychology,* 1939, **8,** 351–379.

BLOOM, B. S. *Stability and change in human characteristics.* New York: Wiley, 1964.

DEARBORN, W. F., ROTHNEY, J. W. M., & SHUTTLEWORTH, F. K. Data on the growth of public school children. *Monographs of the Society for Research in Child Development,* 1938, No. 14.

(Received September 13, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 3, 128–132

CONDITIONS OF PREQUESTIONING AND RETENTION OF MEANINGFUL MATERIAL[1]

D. E. BERLYNE
University of Toronto

Exposure to 28 quotations and subsequently to names represented as possible authors was used with 8 classes of high school students, numbering 225 in all. Retention of the names was significantly improved by requiring Ss to record guesses when 1st exposed to the quotations and by presenting all the authors' names after all the quotations rather than each name immediately after the corresponding quotation. Expressed curiosity about the alleged authors was significantly correlated with retention scores and retention scores with intelligence-test scores, but curiosity was uncorrelated with intelligence. Theoretical implications are discussed.

An experiment carried out some years ago (Berlyne, 1954a) showed retention of factual statements to be enhanced by prior exposure to questions that the statements answered. This finding had been predicted from some hypotheses about epistemic curiosity (Berlyne, 1954b, 1960, 1963, 1965a, 1965b), conceived as a state of heightened drive or arousal that is relieved by acquisition of knowledge. It had been postulated that the questions would generate epistemic curiosity, which would be relieved after subsequent exposure to the corresponding statements and internal rehearsal of them, and that reinforcement from the consequent curiosity reduction would increase the likelihood of recall when the questions were presented again during the test phase. In addition, it was found that the subjects (*S*s) exposed to the forequestionnaire expressed more "extended curiosity" than control *S*s; that is, when asked at the end of the experiment to indicate which of the animals figuring in the material they would like to know more about, they checked off more animals.

Rothkopf (1965) has embarked on a rather similar line of research. He has ascertained that the presentation of questions at various points in the learning process can not only increase retention of facts answering the questions, but also produce a generalized improvement manifested in the retention of other facts. Rothkopf, in interpreting his findings, uses a rather different theoretical approach, holding that questions give rise to "inspective behavior" or "mathemagenic responses," which facilitate meaningful learning.

The experiment to be reported here was meant to carry further in two directions the study of the effects of prequestioning on retention and on extended curiosity. First, effects of three variables relating to conditions of prequestioning were studied. The variables were (*a*) whether or not *S* has to guess the answer to each question, (*b*) time elapsing between consecutive questions, (*c*) whether *S* is exposed to all the questions first and then to all the answers or whether the

---

[1] This investigation was supported by Research Grant MH-06324 from the National Institute of Mental Health, United States Public Health Service. The author wishes to acknowledge the able collaboration of M. A. Craw and J. L. Lewis, who acted as experimenters. He would also like to thank A. B. Patterson, principal of the David and Mary Thomson Collegiate Institute, Scarborough, Ontario, who kindly arranged for his students to serve as subjects.

corresponding answer is given after each question. Second, the experiment was designed to probe effects on retention of two characteristics of multiple-choice questions that had been shown by a previous experiment (Berlyne, 1962) to affect the degree of subjective uncertainty and the degree of reported curiosity. These were (*a*) the number of alternative answers, and (*b*) the degree to which tendencies to select different answers differ in strength.

METHOD

Stimulus Material

The stimulus material, identical with that used in the experiment just mentioned Berlyne, 1962), consisted of 28 brief quotations from men prominent in English or American history or literature of the last 200 years. The quotations were mimeographed in a booklet, and under each quotation were the names of two or three men, ostensibly possible authors. Actually, the name of the correct author was not among them, so that possible effects of previous knowledge could be excluded. Coupled with each name was a number, purporting to be the number of high school teachers out of 100 who had guessed it to be the name of the true author. Of the 28 quotations, seven had three names and a highly uneven distribution of fictitious teachers' guesses, seven had two names and an even distribution, seven had three names and an even distribution, seven had two names and an uneven distribution. These four classes of questions were randomly interspersed.

When the booklets were given to *S*s, approximately equal numbers received each of two versions. The quotations associated with three names in one version were associated with two names in the other and vice versa. The quotations associated with uneven distributions in one version were associated with even distributions in the other, and vice versa.

One of the possible authors associated with each quotation was arbitrarily chosen to be presented as the "correct" author. At the end of the booklet, there was an answer sheet, bearing lines numbered from 1–28, followed by a sheet for the Extended Curiosity Test, bearing the names of the 28 "correct" authors.

Subjects and Experimental Design

*S*s, 225 in all, consisted of four Grade 12 and four Grade 11 classes. The school *S*s attended practiced "streaming" or "homogeneous grouping," so that the A, B, C, and D "forms" in each grade represented different levels of ability.

A 2 × 2 × 2 between-*S*s design was used, the three dimensions being (*a*) Guess versus No-Guess instructions, (*b*) Immediate versus Delayed provision of answers, and (*c*) a 10-second versus a 20-second interval between consecutive questions. The assignment of combinations of treatments to classes is shown in Table 1. It will be seen that the difference between ability levels is confounded with the double interactions and the difference between grades is confounded with the triple interaction.

Procedure

After the experimenter (*E*) had been introduced and *S*s had filled out biographical information on the cover of the booklet, they were told:

> You are being asked to-day to participate in a research project that is being

TABLE 1

ALLOTMENT OF TREATMENTS

Grades	Forms			
	A	B	C	D
11	No guess Immediate 10 sec.	Guess Delayed 10 sec.	Guess Immediate 20 sec.	No guess Delayed 20 sec.
12	Guess Delayed 20 sec.	No guess Immediate 20 sec.	No guess Delayed 10 sec.	Guess Immediate 10 sec.

> conducted by the Psychology Department of the University of Toronto. It concerns ways to improve teaching methods. This experiment is concerned with ability to recognize authors of quotations. As you know, experts can often tell who wrote something from what they know about the background of various authors or from the style of the quotation itself. But at times even an expert will be surprised to find out who the true author is.
>
> Before you open your booklets, let me tell you a little bit about this project:
>
> Some time ago we presented 100 English and history teachers with a number of quotations. Two or three possible authors were listed after each quotation, and the teachers were required to choose which person they thought was the correct author.
>
> You people are going to be shown some of the quotations that were given to the teachers, together with the names of two or three possible authors for each. You will also be given information about the teachers' choices. Printed opposite the name of each possible author will be the number of teachers who chose it as the true author.

The procedure to be followed was then described.

E read aloud each quotation twice while *S*s were required to follow it on the booklet. There was then a pause, during which *S*s read the names of the possible authors and the numbers of teachers' guesses from the booklet. In Guess groups, they were also to underline the name they thought most likely to be that of the true author, but this instruction was omitted for the No-Guess groups. In Delayed groups, all the quotations were treated in this way, and then *S*s turned back to the first page and followed while *E* read aloud the quotations again in the same order, stating the name of the "correct" author after each one. In Immediate groups, *E* read the quotation a third time, followed by the name of the "correct" author, immediately after the pause and before going on to the next quotation. In 20-second groups, the interval between successive quotations in the Delayed condition or between exposure to a quotation and hearing it read a third time with the name of the "correct" author in the Immediate condition was 20 seconds, and in the 10-second groups it was 10 seconds.

After the last "correct" author had been given, *S*s heard *E* read the quotations, in a different randomized order, a fourth time and had to attempt to guess the name of the "correct" author for each one, writing it down on the appropriate numbered line of an answer sheet. Finally, *S*s were required to turn to the Extended Curiosity Test sheet and to check the names of the authors that they would like to know more about.

RESULTS

Retention Scores

The mean Retention scores for the experimental conditions are shown in Table 2. These scores were subjected to analysis of variance and also to analysis of covariance with intelligence-test scores, derived from the Dominion Learning Capacity Test, as concomitant variable. Data from seven *S*s, for whom intelligence-test scores were not available, were excluded from the analysis of covariance.

The within-group correlation between Retention scores and intelligence-test scores was .22, which is significant at the .01 level. The Delayed condition produced significantly better retention than the Immediate condition: $F(1, 200) = 6.05$, $p < .05$ in the analysis of covariance, and

TABLE 2

MEAN RETENTION AND EXTENDED-CURIOSITY SCORES (WITH STANDARD DEVIATIONS IN PARENTHESES)

Mean scores	Guess	No guess	Immediate	Delayed	10 sec.	20 sec.
Retention	9.5 (4.83)	8.0 (4.51)	7.0 (5.10)	10.4 (5.11)	8.5 (4.72)	9.2 (4.71)
Extended curiosity	7.1 (3.55)	6.3 (3.58)	6.6 (3.66)	6.8 (3.54)	6.5 (3.43)	7.0 (3.72)

$F(1, 217) = 6.61$, $p < .05$ in the analysis of variance. The Guess condition produced significantly better retention than the No-Guess condition: $F(1, 200) = 30.01$, $p < .001$ in the analysis of covariance and $F(1, 217) = 33.33$, $p < .001$ in the analysis of variance. Mean Retention scores were higher with 20-second intervals than with 10-second intervals, but not significantly so.

Extended-Curiosity Scores

Table 2 shows the mean for Extended-Curiosity scores (the data for one *S* being lost). The differences are, in all cases, in the same direction as the Retention scores, but analysis of variance shows none of them to be significant.

Extended-Curiosity scores had within-group correlations of .10 (not significant) with intelligence-test scores and .22 ($p < .01$) with Retention scores. The partial correlation of Extended-Curiosity scores with Retention scores, controlling for intelligence-test scores, came to .20 ($p < .01$).

Within-Ss comparisons. When Retention scores for items with three alternative answers were compared with those for items with two answers and when those for items with even distributions of purported teachers' guesses were compared with those for items with uneven distributions, the means were all very close together. There was therefore no significant difference.

DISCUSSION

If we accept the hypotheses that retention increases with epistemic curiosity and that epistemic curiosity depends on conceptual conflict, the superiority of retention in the Guess condition can be explained by assuming that an instruction to guess increases the absolute strengths of all the complexity response tendencies and thus exacerbates competition among them. The more pronounced difference found between the Delayed and Immediate conditions is more surprising and harder to interpret. In the Delayed condition, there is a greater time interval before initial exposure to a particular question and subsequent exposure to its answer. There is also material intervening between the two events, in the form of questions coming later in the series and answers to questions coming earlier, whereas, in the Immediate condition, there is none. It is possible that there is some kind of "incubation" effect: Epistemic curiosity may rise progressively the longer *S* has to wait for the curiosity-relieving information, so that the later the information is received the more curiosity it will reduce. Presumably, if this progressive increase occurs, it will develop only for a certain time and then give way to a decline. There may, however, be other explanations. Research devoted to effects of varying time intervals between exposure to questions and receipt of answers and of varying quantity and nature of intervening experiences would seem to be called for.

The significant positive correlation between Extended Curiosity and Retention could be explained in either of two ways. One possibility is that higher extended curiosity is produced by better retention, because the experience of recalling answers to questions with relative ease makes one favorably disposed to the subject matter or reinforces epistemic (knowledge-seeking) behavior directed toward the subject matter. If this explanation were valid, one would expect any factor that increases retention to increase extended curiosity, and these factors would include the Delayed and Guess treatments. They would also include intelligence, which was positively correlated with retention. However, the differences in Extended Curiosity due to these treatments, while in the same direction as Retention scores, were not significant. Nor was the correlation between extended curiosity and intelligence.

The alternative explanation is that retention depends on at least two distinct factors, among others, one being intelligence and the other being some kind of motivational disposition (cf. the individual differences in curiosity among school children studied by Maw & Maw, 1961). Extended curiosity would then reflect the latter but not the former. The significant partial correlation between Extended Curiosity and Retention, with intelligence kept constant, fits this explanation.

The absence of significant within-*S*s differences in Retention scores due to the two determinants of subjective uncertainty fails to confirm indications from earlier work that retention increases with reported epistemic curiosity. Any effects of these factors are likely to have been weak in comparison with those of other factors (for example, content of questions) so that more sensitive tests of them will be required.

REFERENCES

BERLYNE, D. E. A theory of human curiosity. *British Journal of Psychology,* 1954, **45,** 180–191. (a)

BERLYNE, D. E. An experimental study of human curiosity. *British Journal of Psychology,* 1954, **45,** 256–265. (b)

BERLYNE, D. E. *Conflict, arousal and curiosity.* New York: McGraw-Hill, 1960.

BERLYNE, D. E. Uncertainty and epistemic curiosity. *British Journal of Psychology,* 1962, **53,** 27–34.

BERLYNE, D. E. Motivational problems raised by exploratory and epistemic behavior. In S. Koch (Ed.), *Psychology—a study of a science.* Vol. 5. New York: McGraw-Hill, 1963. Pp. 284–364.

BERLYNE, D. E. Curiosity and education. In J. Krumboltz (Ed.), *Learning and the educational process.* Chicago: Rand McNally, 1965. (a)

BERLYNE, D. E. *Structure and direction in thinking.* New York: Wiley, 1965. (b)

MAW, W. H., & MAW, E. W. Establishing criterion groups for evaluating measures of curiosity. *Journal of Experimental Education,* 1961, **29,** 299–305.

ROTHKOPF, E. Z. Some theoretical and experimental approaches to problems in written instruction. In J. Krumboltz (Ed.), *Learning and the educational process.* Chicago: Rand McNally, 1965. Pp. 193–221.

(Received April 5, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 3, 133–139

REVERSALS IN CHILDREN AS A FUNCTION OF MIDLINE BODY ORIENTATION[1]

ROBERT W. ZASLOW
San Jose State College

The hypothesis that fixed arm movements with respect to the body midline are important factors in written reversals in children was evaluated with 197 normal and 11 cerebral palsied children selected by their teachers as having reversal problems with respect to numbers and letters. A crossover procedure, in which the normal writing position of S was altered by forcing S to write on the opposite side of the body midline, was successful in correcting reversals in 63% of the normal reversal Ss and 79% of the brain-damaged Ss on the 1st trial. In a control group of 30 normal reversal Ss, instructed to correct reversals in the normal writing position, only 13% corrected ($p < .001$).

The reversal of letters and numbers is a constant instructional concern of the teachers of school-age children. Research involving the orientation of figures in space indicates that the proper orientation of a figure is generally not appreciated by young children until 5 years of age (Lord, 1941). As a consequence, the rather high frequency of reversals of letters and numbers in the first and second grades is generally considered to be the result of immaturity. However, it should be emphasized that a great many reversals in children do not seem to reflect error in orientation. A variety of other theories have been presented to explain reversals in children, especially reversals which persist through and beyond the primary grades. Among these theories are handedness, in which a child is forced to write with a nondominant hand or demonstrates a confusion because of a lack of hand dominance (Travis & Johnson, 1934). Closely allied to this position is the proposal that reversals are produced as a result of mixed cerebral dominance (Travis & Johnson, 1934). Other investigations attribute reversals to central nervous system damage in those children with brain injury who seem to show a high incidence of reversals. Perceptual and figure-ground disturbance are said to account for reversals in brain-injured children. Emotional disturbance is also posed as a causative factor in children's reversals. The Bender Visual Motor Gestalt Test indicates that reversals and rotation of design are frequently found in both brain-damaged and emotionally disturbed subjects (Koppitz, 1964; Pascal & Suttell, 1951).

Of these theories only those involving emotional disturbance, where perceptual and spatial disorganization may occur, and central nervous system damage, which also produces perceptual and spatial disturbance, have been adequate in certain cases of reversals. Immaturity may not account for persistent reversals after the age of 6, except perhaps in a small proportion of children. The writer feels that the greater incidence of reversals found in boys, usually attributed to immaturity, may be due to greater motor resistance arising from anxiety or from inhibition of aggression.

The importance of the body midline in understanding reversals in children has been neglected, and the writer suggests fixed motor responses as a major etiological factor in persistent reversal cases. These motor responses appear to be based on primitive behavior patterns and are, to a large degree, independent of volitional or cognitive control. Emotional factors may reinforce the rigidity of these motor patterns.

---

[1] This study was sponsored by a grant from the San Jose State College Faculty Research Fund.

From this point of view, it follows that for some children reversals should be corrected on the motor level rather than on cognitive or emotional levels exclusively. Radler and Kephart (1960) refer to the body midline as a factor in reversals and orientation problems only tangentially and do not extract the full significance of motor movements in writing with respect to the body midline. The fact that more boys have persistent reversal problems than girls and show greater muscular aggressiveness, thrust, activity, and emotional resistance supports the position of the writer.

The purpose of this study is to suggest the importance of the body-midline function in a method for correcting reversals in normal, retarded, and brain-damaged children. This method follows from the above theoretical considerations about hand motor patterns with respect to writing movements to and away from the vertical midline body axis, and the suggestion that certain primitive motor patterns become fixed with respect to the body midline and may persist independently of the child's intention to produce a correctly orientated figure. Thus, certain motor patterns may be responsible for the production of persistent reversals that are resistant to change by the usual instructional methods.

In experimenting with a number of reversal cases some time ago, the investigator observed that often a letter or number that had been reversed with either hand was immediately corrected if the writing hand was positioned so that it crossed the body midline. In other words, if the right hand produced a reversal in normal writing position on the right side of the body, the reversal became corrected if the right hand was placed on the left side of the body midline without moving the body. This occurs because the motor pattern now is a mirror image of the reversal in the normal position. It is self-evident that a motor pattern going away from the midline in the normal writing position will become reversed and go toward the midline when the writing hand is positioned on the opposite side of the midline.

This technique was used in a reversal study involving first-, second-, and third-grade students, as well as a small number of brain-damaged children. These children were evaluated with respect to possible correction of reversals using the midline-crossover procedure.

METHOD

Subjects

Total subjects were 208 children, all evaluated using the crossover procedure, who were selected by their teachers as having reversal problems with respect to numbers and letters. The original experimental group consisted of 178 children (167 normal and 11 ungraded cerebral-palsy children); the control group consisted of 30 normal children. In the original experimental group, 52% were age 6, 38% were age 7 (the cerebral-palsy children accounted for the higher age range); 64% were from Grade 1, 27% from Grade 2 (the cerebral-palsy children were ungraded); 61% were boys, 39% were girls. Of the original experimental group of 178, only 80 (45%) were found to actually have reversal problems in an individual testing situation, leaving a final total of 80 subjects in the experimental group (including the cerebral-palsy children) and the control group consisting of 30 normal subjects. Of the 80 subjects in the final experimental group, 53% were age 6, 35% were age 7; 61% were from Grade 1, 25% from

Grade 2; 60% were boys, 40% were girls. In the control group 68% were age 6, 30% were age 7; 59% were boys, 41% were girls (see Table 1).

Procedure

Each child was given a large sheet of white paper divided down the center by a line representing the midline of the child's body. Then he was requested to write the numbers and letters which had been found to present reversal problems. These numbers and letters were to be written in a descending vertical order along the midline, on the right side of the midline for right-handed children and on the left side of the midline for left-handed children. This step was taken to determine if the child would correct the reversal without crossing the midline. Upon completion of the right-sided set, the right-handed child was told to cross his hand over the midline to the left side and again write those letters and numbers which he had reversed. The examiner took care to prevent possible twisting action of the body which would change the direction of the child's midline. The original experimental design utilized the experimental groups as their own controls. At a later date a separate control group of normal reversal cases was evaluated as an improvement on the original design. The control group was instructed to repeat their reversed letters and numbers three times while maintaining the same side of the normal writing position.

TABLE 1

AGE, GRADE, AND SEX OF *S*s IN FINAL EXPERIMENTAL AND CONTROL GROUPS

Age-grade levels of *S*s	Experimental group		Control group	
	No. *S*s	% *S*s	No. *S*s	% *S*s
Age				
6	42	53	20	68
7	28	35	9	30
8	6	8	1	2
9	2	2	—	—
10–17	2	2	—	—
Grade				
1	54	61	20	68
2	20	25	9	30
3	—	—	1	2
Ungraded	6	14	—	—
Sex				
M	48	60	18	59
F	32	40	12	41

TABLE 2

FREQUENCY OF REVERSALS AND CORRECTED REVERSALS

Categories of Analysis	Reversal frequency per *S*					Total no. *S*s
	1	2	3	4	5	
% *S*s in each frequency category						
Experimental group	44	32	11	7	6	80
Control group	53	33	13	6	3	30
% *S*s correcting in relation to total number *S*s correcting						
Experimental group	49	26	14	9	2	51
Control group	50	25	25	—	—	4
% *S*s correcting in relation to total number *S*s in each category						
Experimental group	71	52	78	83	20	
Control group	12	10	10	—	—	

RESULTS

Experimental Group

It was found that many of the children selected for the study did not demonstrate the reversals that they usually produced in a classroom situation. In a face-to-face individual testing situation only 45% of these subjects reversed symbols. Inattention factors may be operative in the classroom and greatly reduced in the one-to-one testing situation. As expected, there was a sharp decrease of reversals in the higher age and grade levels. The majority of reversals were found in boys (60%); this result is in agreement with, and to be anticipated from, the literature.

As Table 2 indicates, the majority of all subjects in the experimental group reversed only one or two symbols ($M = 2.1$). The results show that 60% of all the reversals produced by the experimental subjects were corrected by the midline method. When the total number of subjects correcting reversals by the midline method is considered, 63% of the subjects in the experimental group corrected *on the first trial* (see Table 3).

Results show that in the brain-damaged (cerebral-palsy) children, reversals were distributed throughout the various grade and age levels, and a larger proportion (72%) of this group reversed than in the total population. It is in this group that we would expect to find more of the fixed motor patterns that could be a causative factor in reversals, and it is also in this group that the greater percentage of corrections (79%) were shown. These results hold promise and should be investigated more extensively in brain-damaged children.

The most frequently reversed numbers, in order of occurrence, were 2, 7, 3, and 9. Only 9% reversed 6, 8% reversed 5, and 6% reversed 4. The number most frequently corrected was 2 (25%), closely followed by 7 (19%) and 3 (17%). We may note that numbers 2, 7, and 3 require an initial movement *away* from the midline so that it may very well be difficult for some subjects to initiate an arm movement away from the midline, while other subjects have trouble initiating movements toward the midline as in numbers 6 and 9.

The most frequently reversed letter was Z (13% of the cases), closely followed by b (12%) and C (10%). Although certain reversals were more common, there seemed to be a wide distribution of reversals among all letters, and the corrected letter reversals also fell in a wide distribution. The most frequently corrected letter was C (12%), while B, D, Q, Z, b, and d were each corrected by 7% of the subjects. It may be noted that Z reverses its direction *toward* and *away* from the midline, which may account for the high frequency of reversals. In most instances reversals appear to result from a fixed motor set which initiates movements toward the midline.

TABLE 3

TOTAL REVERSALS AND CORRECTED REVERSALS

Categories of Analysis	Experimental group	Control group
Total no. *Ss*	80	30
% *Ss* correcting reversals	63	13
Total no. reversals	161	60
% reversals corrected	60	10

Control Group

In contrast to the experimental group, the control group showed a total of only 10% correction of reversals on the first trial. The average number of reversed items was 2.0, and the total number of children correcting reversals was 4 (13%). The effectiveness of the crossover method is indicated by the superior results of the experimental group in correcting reversals. The chi-square test showed the significance to be at the .001 level.

DISCUSSION AND CONCLUSIONS

A suggested method for correcting reversals is as follows: Have the right-handed child cross the midline with the right hand and write the symbol which he reverses. If the symbol is corrected by this procedure, then, while the hand is still crossed over the midline, the child should make a row of the symbol, moving toward the right and crossing the midline, so that the end of the row falls into the proper writing position on the right side of the midline.

This midline-crossover technique

appears to be fairly successful in correcting a large number of reversals in normal and brain-damaged children on the first trial. While it is quite probable that reversals may result from several factors, this technique appears to be effective in correcting a core factor in the production of reversals—namely, movements to or away from the midline. The importance of the midline function may be elaborated upon to clarify and aid in certain instruction problems that the teacher encounters. It has been noted in this study that certain children are capable of producing normal numbers and letters only if they begin to make them from the bottom rather than from the top. This procedure compensates for certain fixed motor patterns which produce reversals if the child is fully guided by them. For example, if a 7 is started at the bottom rather than the top, a tendency to make a movement toward the midline will not produce the reversal that would occur if it were written in the normal manner.

The brain-damaged (cerebral-palsy) subjects produced more reversals and obtained a higher percentage of corrected reversals than the normal subjects. This result is to be expected because of the fixed motor patterns resulting from cortical insult. This study suggests that fixed motor patterns play a major role in the symbol reversals of cerebral palsied children; perceptual disturbances may not be the major cause of reversals in these subjects. It is possible that rotations on the Bender Gestalt Test, so characteristic of these subjects, may result from improper placement of the protocol sheet with respect to the midline of the subject. The subject may orient designs according to his own midline as projected on the sheet of paper instead of the actual position of the paper. This may occur when the paper is placed at an angle to the body midline.

Left handedness complicates the entire problem of midline orientation and may account for the greater number of reversals found in left-handed children. The left-handed child may position the writing hand across the midline in a manner and direction that increase the complexity of the writing motor act with respect to midline orientation, thereby increasing the possibility of left-right type reversals of symbols. In addition to the mere mechanics of writing with the left hand, there is an emotional and personality factor. The literature suggests that left handedness itself is a symptom of negativism and oppositional behavior at an early age. Toilet training produces more aggression and resistance in boys (Sears, Whiting, Nowles, & Sears, 1953). Left handedness, stuttering, and reversals, more common in boys, may be linked to this period of childhood through the common bond of motor resistance resulting from emotional dynamics, usually anxiety and hostility. These early and primitive motor patterns may be quite resistant to change. A child may resist with arms and hands in such a manner that he is not aware of the resistance and, as a consequence, has neither guilt nor control over his reversal tendency in writing. Thus, the fixed motor pattern in question may have an emotional as well as a motor conditioning basis. Anxiety may either fixate responses or produce excessive vacillation of such responses with respect to midline orientation.

An analysis of the literature lends support to the theoretical position underlying this study and reinforces the impression that boys are not inferior to girls with respect to basic skills. Boys are not more immature; they are different. Werner (1940, 1957) postulates that the most primitive concepts of spatial relationships involve dynamic visual movement (such as in a tapping test), while the more highly refined concepts of space are more abstract and impersonal and involve a minimum of, or no, movement. These findings suggest a causal link whereby boys may suffer greater impairment of concepts of spatial relationship because of muscular aggressiveness or hyperactivity as an expression of anger.

Goodenough (1931) found that there is a much greater frequency of anger and aggression in boys than girls after the age of 2 and that this sex difference is maintained up to the age of 7 and beyond. Thompson (1962) remarks that this fairly consistent sex difference in the frequency of anger episodes is interesting and enigmatic at our present state of knowledge. The writer believes that the overt expression of anger in boys is one of the major causative factors in fixing of rigid motor patterns to or away from the body midline.

Michael, Zimmerman, and Guilford's (1951) experimental work provides evidence that spatial relations continue to be coordinated to the human body even into the elementary school years. They also found that a factor analysis of 15 different tests of spatial relations and visualization revealed definite superiority in the boys. Hildreth (1932), in her excellent monograph on the subject of reversals, contends that the greater incidence of reversals in boys is a logical consequence of the learning process and an indicator of immaturity, an argument that is not supported by the findings that show boys to be superior in motor ability and spatial relations.

The developmental importance of the body midline to writing has been suggested by Ames (1948). She found that children, when writing on a sheet of paper, tend to move from the lower right-hand corner of the paper up to the center and then up to the top center, and that these orientation changes are sequential and orderly.

Speed of reaction and activity levels in boys may account for reversals. Koch's (1934) study, which involved a factor analysis of activity in nursery school children, found that speed of reaction is often classified as nervousness or aggressiveness by observers. This speed may actually be inhibited by the classroom teacher and produce motor resistance. The writer also feels that the superior speed of response found in boys could produce more spontaneous primitive motor responses to or away from the midline than would appear when reactions are more deliberate. Pomeroy (1938), studying children 5½-years-old, found that extroverted boys are speedier than introverted boys, while girls fell in the middle range of the extroversion-introversion scale.

The writer feels that these studies generally support the theoretical position taken with regards to the significance of the body-midline axis in the production of reversals in children and points to the need for the investigation of this phenomena for the purpose of motor retraining as well as theory construction. On the basis of this research, primarily a pilot study, it would be profitable to enlarge the study to include a larger sample of brain-damaged subjects as well as mentally retarded children. Additional research is required to determine whether this technique can facilitate the instruction of number and letter writing in the classroom and permit the permanent correction of reversals in persistent reversal cases.

REFERENCES

AMES, L. B. Postural and placement orientations in writing and block behavior: Development trends from infancy to age ten. *Journal of Genetic Psychology,* 1948, **73,** 45–52.

GOODENOUGH, F. L. *Anger in young children.* Minneapolis: University of Minnesota Press, 1931.

HILDRETH, G. The success of young children in number and letter construction. *Child Development,* 1932, **3,** 1–14.

KOCH, H. L. A multiple-factor analysis of certain measures of activeness in nursery school children. *Journal of Genetic Psychology,* 1934, **45,** 482–487.

KOPPITZ, E. M. *The Bender Gestalt Test for young children.* New York: Grune & Stratton, 1964.

LORD, F. E. A study of spatial orientation of children. *Journal of Educational Research,* 1941, **34,** 481–505.

MICHAEL, W. B., ZIMMERMAN, W. S., & GUILFORD, J. P. An investigation of the nature of the spatial-relations and visualization factors in two school samples. *Educational and Psychological Measurement,* 1951, **11,** 561–577.

PASCAL, G. R., & SUTTELL, B. J. *The Bender-Gestalt Test; qualification and validity for adults.* New York: Grune & Stratton, 1951.

POMEROY, J. E. The relation of reaction time of five-year-old children to various factors. *Child Development,* 1938, **9,** 281–283.

RADLER, D. H., & KEPHART, N. C. *Success through play.* New York: Harper, 1960.

SEARS, R. R., WHITING, J. W. M., NOWLES, V., & SEARS, P. S. Some child-rearing antecedents of aggression and dependency in young children. *Genetic Psychology Monographs,* 1953, **47,** 135–236.

THOMPSON, G. G. *Child psychology.* Boston: Houghton Mifflin, 1962.

TRAVIS, L. E., & JOHNSON, W. Stuttering and the concept of handedness. *Psychological Review,* 1934, **41,** 534–562.

WERNER, H. Perception of spatial relationship in mentally deficient children. *Journal of Genetic Psychology,* 1940, **57,** 93–100.

WERNER, H. (Ed.) *Comparative psychology of mental development.* New York: International Universities Press, 1957.

(Received May 18, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 3, 140–145

EFFECTS OF DELAY OF INFORMATION FEEDBACK AND LENGTH OF POSTFEEDBACK INTERVAL ON LINEAR PROGRAMED LEARNING[1]

FREDERICK J. BOERSMA[2]
Department of Educational Psychology, University of Alberta

The study attempted to determine the independent effects of delay of information feedback (IF), post-IF delay, and sex in a complex learning task in which Ss learned a series of symbolic logic rules by programed instruction. 56 Ss served individually in a factorial design which combined 2 delays of IF (0 and 8 sec.), 2 post-IF intervals (0 and 8 sec.), and a sex factor. The data did not support the prediction that human learning would be facilitated by increasing the post-IF interval, nor did it yield a significant delay of IF effect. But it did show that a simple explanation in terms of the independent effects of delay of IF or post-IF is inadequate. A competing-response interpretation of delay was formulated with respect to error scores, answer latencies, and the delay variables.

Though the study of experimental delay goes back nearly 50 years, and though research has demonstrated the retardation effects of delay on lower organisms, it has failed to generate consistent findings with humans (Renner, 1964). Thus, the complete acceptance, without qualification, of delayed-reinforcement or information-feedback (IF) generalizations, which suggest that a few seconds delay between response and reinforcement may mean the difference between maximal learning and no learning whatever, is unwarranted in human learning.

Bourne and Bunderson (1963) have suggested that the controversial delay effects in human learning may be due to the fact that experimenters have failed to control for confounding effects of the intertrial interval (ITI) with post-IF delay and have failed to consider methodological factors as task complexity. To control for such confounding effects, experimenters developed a design which permitted them to vary these variables orthogonally while maintaining an average ITI between treatments. They found that delay of IF did not retard concept attainment, but that increases in the post-IF interval led to significant decreases in error scores, and the post-IF effect became more prevalent as concept complexity increased.

Brackbill and Kappy (1962), however, have suggested that discrepancies may be attributable to the fact that the subjects (*S*s) can mediate the delay interval in some cases, but not in others. Yvonne Brackbill (1964, unpublished)[3] provided some support for this hypothesis in a serial learning study. She found that an interpolated task initiated during delay retarded performance and concluded that decremental effects associated with delay increase when *S*s cannot reinstate symbolically some form of the response aftereffect. Her study did not, however, control for confounding effects of the ITI with post-IF delay, nor did it consider task complexity.

---

[1] This paper is based on a dissertation of the same title submitted to the Graduate College of Michigan State University in partial fulfillment of the requirements for a PhD degree.

[2] The author is indebted to Bernard R. Corman for his valuable assistance and advice in the conducting of this study.

[3] The manuscript, "The Impairment of Learning under Immediate Reinforcement," can be obtained in dittoed form from its author at the University of Colorado Medical Center.

Thus, it appears that variables other than delay of IF per se must be considered when making generalizations regarding the effects of delayed reinforcement on human learning, and that typical reinforcement generalizations based on lower organisms and simple tasks may not be applicable to actual school learning. Results from studies employing complex tasks more closely related to school learning, as in programed instruction, which control for confounding effects of delay with the ITI are now needed.

The primary purpose of the present study was to extend the work of Bourne and Bunderson (1963) by using a task closely resembling actual school learning, a programed-learning task in symbolic logic. A secondary purpose was to examine verbal reports (open-ended questionnaire responses) obtained from *S*s at the conclusion of the learning task. It was surmised that such data would provide valuable information concerning the influence of delayed reinforcement on a complex learning task and shed light on mediational processes employed by *S*s during delay intervals.

The study differs from traditionally reported learning research on delay in that the differential complexity of material within the program necessitated the use of subject- rather than experimenter-controlled feedback intervals. But this change in experimental procedure was imperative if *S*s were to receive adequate IF. It was predicted that all correlations between error scores and IF intervals for the respective treatments would be negligible, and that consequently this new variable would not confound the study. The hypothesis was supported.

Figure 1 represents the temporal intervals common to this study. The ITI was defined as that period of time between response and the presentation of the next stimulus. The delay of IF interval referred to the amount of time *S*s waited after responding before receiving IF, while the post-IF interval referred to the amount of time *S*s waited after receiving IF before being exposed to the next frame. The IF interval was synonomous with the answer-latency variable.

It was predicted that the post-IF interval would be the crucial variable in this complex learning task, and that the effect of delayed IF would be negligible. It was further surmised that the data would resemble Bourne and Bunderson's (1963), since for all practical purposes the experimenter (*E*) had controlled for confounding effects of ITI with the delay variables and had used a difficult (complex) learning task.

PROCEDURE

Subjects

Fifty-six volunteer educational psychology students, 28 men and 28 women, served individually as *S*s in this experiment. Their median age was 20 years, and their cumulative grade-point average, 2.66. These *S*s were randomly assigned to the four experimental conditions (14 per experimental treatment) with sexes equally represented in each treatment.

Task and Apparatus

The linear program in symbolic logic developed by Evans, Glaser, and Homme (1962) was modified for use in this experi-

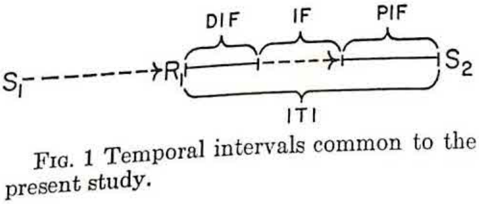


FIG. 1 Temporal intervals common to the present study.

ment. The final form of the modified program consisted of 7 rules to be learned in 78 frames. Each frame was programed according to the Ruleg System (Evans, Homme, & Glaser, 1962).

The apparatus consisted of an MTA Scholar teaching machine, a Hunter timer, and an Esterline-Angus recorder. The teaching machine and timer were wired so that *E* could control the delay intervals. The Esterline-Angus recorded latency to response and amount of time *S*s spent examining the correct answer. The program was advanced by the internal mechanism of the apparatus when *S*s pressed a button.

*S*s were allowed to control the amount of time spent on each frame. After *S*s constructed a response, they were informed of the correct answer either immediately or after an 8-second delay. The amount of time spent examining the correct answer was also controlled by *S*s. After this second self-determined interval, they were exposed, either immediately or after an 8-second delay, to the next frame. In all treatments, the stimulus material and *S*'s constructed response were withdrawn from view during the delay intervals. The lengths of the respective delays, that is, the delays of IF and post-IF, were predetermined as a function of the treatment conditions.

Each *S* received a 5-minute break midway through the program and before administration of a criterion test after completion of the program. This test consisted of 10 true-false questions, 10 recall questions, and 8 deductive-proof questions; an open-ended questionnaire was also included in the battery. Average total-program running time per *S* was approximately 2 hours, and varied from 1 hour, 25 minutes to 2 hours, 35 minutes.

The dependent variables were two error scores, a response-latency score, and an answer-latency score. Number of errors indicated to what degree *S*s responded correctly to program and criterion-test items. Response latency referred to the average amount of time spent by *S*s responding to programed items, and answer latency to the average amount of time spent by *S*s examining the corect answers to these items.

Scorer reliability (test-retest) over a 7-day period was .963 for program errors and .981 for criterion-test error scores. The reliability of the criterion test was .843 (KR 21), whereas the correlation between program and criterion error scores was .875. Recorder reliability (interjudge) was .994 for response latencies and .998 for answer latencies.

Design

A 2 × 2 × 2 factorial design was used with 2 delays of IF (0 and 8 seconds), 2 lengths of post-IF interval (0 and 8 seconds), and a sex factor. The experimental design permitted *E* to vary the effects of delay of IF and post-IF orthogonally, to maintain an average intertrial interval of 8 seconds between these 2 factors, and to control for possible massing effects between the various treatment conditions.

The 4 treatment conditions were identified as follows: 0-0 indicated complete self-pacing by *S*s (i.e., no delay of IF or post-IF delay); 0-8 indicated no delay of IF, but an 8-second post-IF interval; 8-0 indicated an 8-second delay of IF, but no post-IF delay; and 8-8 signified an 8-second delay interval before and after exposure to correct answer (i.e., before and after IF).

RESULTS AND DISCUSSION

The analysis of variance on program error scores revealed that all main and interaction effects were insignificant except the delay of IF × post-IF interaction ($F = 5.640$, $df = 1/48$, $p < .05$). Mean number of errors for Treatment 0-0 was 7.62; for Treatment 0-8, 11.62; for Treatment 8-0, 13.30; and for Treatment 8-8, 9.52. Similar, though not as pronounced, findings were obtained for the analysis-of-criterion error scores. In this analysis, the delay of IF × post-IF interaction failed to reach the .05 level of significance; however, it did exceed the .10 tabled value of the *F* distribution ($F = 2.821$, $df = 1/48$). Figure 2 represents graphically the delay of IF × post-IF interactions for program and criterion error scores.

A significant sex effect was found for the analysis of average response latencies indicating that females spent less time responding to each frame on the average than did males ($F = 4.085$, $df = 1/48$, $p < .05$). All other main and interaction effects for response latencies were insignificant. A similar analysis was performed on average program answer-latency scores. This breakdown revealed, as did the previous analysis, that female latencies were on the average shorter than those of males ($F = 4.908$, $df = 1/48$, $p < .05$). And, more important, it revealed that *S*s spent a significantly greater amount of time examining correct answers when their treatments were associated with delayed, rather than immediate, feedback conditions ($F = 12.676$, $df = 1/48$, $p < .01$).

*S*s' responses on the open-ended questionnaire were independently categorized by two judges into behavioral classifications indicative of fatigue (physical) and frustration (competing responses). Representative examples of fatigue responses are: "It seemed to get tedious towards the end"; "I felt somewhat discouraged and became aware of my tiredness towards the end"; and "I kept thinking, 'When will it ever be finished?' " Examples of frustration responses are: "I wished the machine would hurry up—it was frustrating when you had to wait"; "It was somewhat confusing as the rules multiplied"; and "I found myself wishing I had a reference to all rules in relation to each other at one time for study." There was 95% agreement between the independent classifications of the two judges. Table 1 gives percentage breakdowns of open-ended questionnaire responses as a function of treatment.

TABLE 1

Percentage of Frequency of Reported Fatigue and Frustration Responses on Open-Ended Questionnaire as a Function of Treatment

Treatment		Type of response	
DIF	Post-IF	Fatigue	Frustration
0	0	43[a]	0
0	8	0	57
8	0	7	57
8	8	7	57

[a] $N = 14$ per cell.

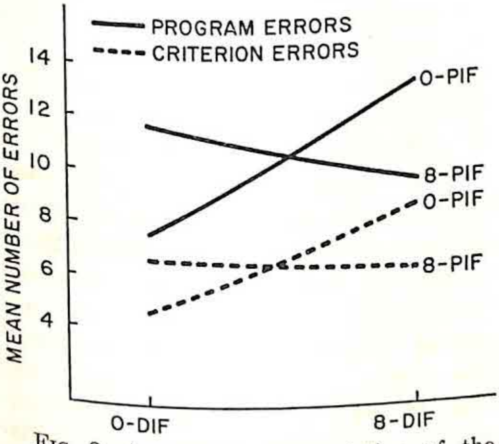


Fig. 2. A graphic representation of the delay of IF × post-IF interaction for program and criterion-test error scores.

The present study fails to support Bourne and Bunderson's (1963) supposition that the post-IF interval is a critical variable in human learning. Where Bourne and Bunderson found that increases in the post-IF interval were accompanied by decreases in error scores, the effects of post-IF delay in this experiment were negligible. But the position of Bourne and Bunderson that delay of IF is not a critical variable in complex human learning was supported.

Both in Bourne and Bunderson's and in the present study, delay of IF and post-IF delay were varied orthogonally. The essential difference between these two studies was that Bourne and Bunderson's tasks were peaked at a constant level of difficulty such that IF could be presented in a fraction of a second (a light flashed on the apparatus), and, as a result, *E* completely controlled the ITI. In the present study task, however, complexity varied throughout the program and necessitated the use of subject- rather than experimenter-controlled IF, and consequently complete *E* control of the ITI was impossible.

Thus, Bourne and Bunderson's treatments are relatively more massed than are those used in the present study. The effects of massing should be most prevalent in their self-paced treatment, 0-1, and consequently this treatment should be associated with maximum error scores. Massing effects should not occur, however, under self-pacing, such as that used in the present study, where *S* controls the length of the IF interval. Moreover, the rank order between comparable-delay treatment means in Bourne and Bunderson's and the present study should be similar since massing should dissipate when delay of IF and/or post-IF delay is introduced into the experimental situation.

This interpretation was supported by a comparison of comparable treatments between the two studies.[4] Where Bourne and Bunderson found their 0-1 treatment to be associated with the largest number of errors, in the present study the comparable 0-0 treatment produced the fewest number of errors. An identical rank order was found between all treatment means where delay intervals were similar.

Even though massing per se did not appear to be a crucial factor affecting performance in the present study, 43% of *S*s in the self-paced treatment did complain of physical fatigue due to length of program, whereas 7% or less made similar complaints in each of the delay treatments. If fatigue due to massing is operating during complete self-pacing, as it appears to be, its influence should be less than that associated with self-pacing in which *E* provides IF in a fraction of a second.

The answer-latency data indicated that a significantly greater amount of time was spent by *S*s examining correct answers when treatments were associated with delayed- rather than immediate-feedback conditions. It appears that when IF was delayed, the number of interfering (competing) responses which *S* had to process and/or eliminate while examining the correct answer was increased, and consequently it was necessary for him to spend more time examining correct answers under delayed- than immediate-feedback conditions. Support for this interference interpretation was obtained from the analysis of questionnaire responses which revealed that 57% of *S*s in each of the delay treatment complained of frustration (competing responses) while working through the program.

In sum, it is surmised that: (*a*) Competing responses are initiated during delay; (*b*) these responses are overtly characterized in terms of frustrative behavior; (*c*) these responses become conditioned to the same stimuli as does the to-be-learned response, and consequently acquisition time increases as a function of increments in the delay variable; and (*d*) *S*s are able to use the IF interval to process and/or eliminate many of the competing responses elicited during delay.

---

[4] Bourne and Bunderson's five-dimensional concept treatments were used for this comparison. These treatments were chosen since the complexity of these tasks more closely resembled that of the tasks used in the present study, and since Bourne and Bunderson found the post-IF effect to be more pronounced for complex concepts.

Moreover, it appears that if delay of IF had been extended to the point where *S*s could no longer cope with these competing responses, a significant delay of IF effect would have emerged. Similar competing-response interpretations of delay have been formulated by Spence (1960) and Denny and Adelman (1955) for animal learning, and by Denny[5] for human learning.

REFERENCES

BOURNE, L. E., & BUNDERSON, C. V. Effects of delay of informative feedback and length of postfeedback interval on concept identification. *Journal of Experimental Psychology,* 1963, **65,** 1–5.

BRACKBILL, Y., & KAPPY, M. S. Delay of reinforcement and retention. *Journal of Comparative and Physiological Psychology,* 1962, **55,** 14–18.

DENNY, M. R., & ADELMAN, H. M. Elicitation Theory I: An analysis of two typical learning situations. *Psychological Review,* 1955, **62,** 290–296.

EVANS, J. L., GLASER, R., & HOMME, L. E. An investigation of "teaching machine" variables using learning programs in symbolic logic. *Journal of Educational Research,* 1962, **55,** 433–452.

EVANS, J. L., HOMME, L. E., & GLASER, R. The Ruleg System for the construction of programmed verbal learning sequences. *Journal of Educational Research,* 1962, **55,** 515–520.

RENNER, K. E. Delay of reinforcement: A historical review. *Psychological Bulletin,* 1964, **61,** 341–361.

SPENCE, K. W. *Behavior theory and learning.* Englewood Cliffs, N. J.: Prentice-Hall, 1960.

(Received May 20, 1965)

---

[5] M. R. Denny, personal communication, November 1964.

Journal of Educational Psychology
1966, Vol. 57, No. 3, 146–150

EIGHT PROFESSORS—A STUDY ON COLLEGE STUDENTS' PREFERENCES AMONG THEIR TEACHERS

KAORU YAMAMOTO
University of Iowa

AND

HENRY F. DIZNEY
Examination Aids Center,
Kent State University

Students' preferences for different types of college professors, Administrator, Socialite, Teacher, and Researcher, from each of 2 sources, education and arts and sciences, were studied. 87 graduate students, 131 seniors, and 82 juniors in secondary education responded to a Likert-type inventory, the Eight College Professors. Type and Source of professors were significant sources of variation. Teacher, Researcher, Socialite, and Administrator were preferred in that order, and education was preferred to arts and sciences as a source. Professor Type × student Level interaction was also significant. Researcher was the only type to gain in preference from juniors to graduates. No remarkable sex differences were observed.

McGee (1962) has discussed the "disorganized" nature of the American university and attributes that condition to its threefold (creation-conservation - dissemination) function. Others agree in principle if not in terminology (e.g., Knapp, 1962; Pusey, 1963). Shifting importance attached to these functions by various elements of the university are suggested by both Knapp (1962) and Caplow and McGee (1961). Some of the concomitants of such change appear disturbing. Jacob (1957) suggested that the entire function of the teaching profession has fallen into low esteem among students. A more sober picture was drawn by Kerr (1963):

> The undergraduate students are restless. Recent changes in the American university have done them little good.... There is an incipient revolt of undergraduate students against the faculty; the revolt that used to be against the faculty *in loco parentis* is now against the faculty *in absentia* [p. 103].

Faculty members may be in absentia in ways other than physical. Since academic advancement often seems more related to nonteaching than teaching activity, it may be expected that the press of other demands can adversely affect the quality of instruction. On the basis of observation it seems that four rather distinct "types" of faculty personnel exist. For this study they were named Administrator, Socialite, Teacher, and Researcher. Three of the four types tend to be in absentia because of the very nature of their interests.

Evidence indicates that the perceptions of students regarding teachers are more a function of internal frames of reference than of concrete teacher characteristics (Gage, 1953; Katz, 1962). Knapp's (1962) summaries point out that students tend to have sharply defined, consistent images of professors and emphasize personal-social rather than intellectual characteristics in ratings of "effective" teachers.

It may be expected that: (*a*) Students, given a choice, would prefer as their professors Teachers and Socialites to Researchers and Administrators; (*b*) female students would reveal a pattern of larger differential preference than male students; (*c*) advanced students would show a stronger preference for the last two groups to the first two than less-advanced students; and (*d*) professors from liberal arts would be preferred to those from education. These predictions are based either upon prior findings or common observation reflecting prestige hierarchies. Thus, this study was an exploration of student preference for teachers in relationship to (*a*) types of professors, (*b*) students' sex, (*c*) students' educational level, and (*d*) collegiate identification (source) of professors.

METHOD

Subjects

*S*s of the study were 300 secondary education students at Kent State University. Of these, 87 were graduate students, 131 were seniors, and 82 were juniors.

Procedures

*S*s were administered in group sessions the following instrument.

Eight college professors. This was an 8-item, 5-choice, Likert-type inventory constructed by the authors, taking their lead from Phillips (1963, 1964) and Tannenbaum (1962). Four types of professors (Administrator, Socialite, Teacher, Researcher) from two collegiate sources (education and liberal arts) were described in each of eight paragraphs. *S*s rated each description on a continuum from "strongly prefer" (5 points) through "neutral" (3) to "strongly reject" (1). The professor descriptions follow.

Professor A is an administrative assistant to the Director of Admissions. He has always been interested in college administration and aspires to higher office. He holds an academic appointment with the College of Arts and Sciences and teaches (and has taught for 10 years) a course in educational methodology. He has a doctorate from a well-known university, is married and has three children. Besides his classroom contacts, he works with the registrar, deans, treasurer, and other university administrators, as well as administrators from other colleges and universities. [For *Professor E*, change the College of Arts and Sciences to the College of Education.]

Professor B is advisor to the largest social fraternity on campus. He has been an active member of a lodge group for a number of years and aspires to gain greater social prestige. He holds an academic appointment with the College of Education and teaches (and has taught for 10 years) a course in educational methodology. He has a doctorate from a well-known university, is married and has three children. Besides his classroom contacts, he socializes with faculty, administrators, and non-university people on a wide basis. He is a frequent attendant at plays, concerts, parties, and conventions. [For *Professor F*, change the College of Education to the College of Arts and Sciences.]

Professor C is an academic advisor to students. He takes seriously his role as student mentor and devotes much time to the preparation of his lectures. He enjoys teaching and aspires to become a better teacher. He holds an academic appointment with the College of Arts and Sciences and teaches (and has taught for 10 years) a course in educational methodology. He has a doctorate from a well-known university, is married and has three children. Besides his classroom contacts, he works with students and colleagues toward more effective instruction. He spends a great deal of time helping students in their academic work. [For *Professor G*, change the College of Arts and Sciences to the College of Education.]

Professor D directs a large-scale research project. He has always been interested in basic research and aspires to expand his research activity. He holds an academic appointment with the College of Education and teaches (and has taught for 10 years) a course in educational methodology. He has a doctorate from a well-known university, is married and has three children. Besides his classroom contacts, he associates himself with other active researchers. His research results in the presentation of papers at professional conventions and the publication of articles in specialized journals. [For *Professor H*, change the College of Education to the College of Arts and Sciences.]

The descriptions were assembled in random order for each *S* on 8 4½ × 8-inch cards. The instrument was then presented with these instructions:

As you may know, it is a frequent

practice at many colleges to employ instructors for educational methods courses from either Arts and Sciences or Education faculties. It is also well-known that professors, regardless of their specialty, vary widely in terms of personal characteristics. The packet describes eight general types of professors. Read and rate these professors one at a time in the order they appear in the packet, i.e., first appearing professor should be rated *before* reading the second description, etc. For example, your first professor may be any between A and H, but assume that it is D. Read D first and be careful to rate him in D's position on the answer sheet. Are there any questions?

When scores on each type were correlated between sources for *S*s, the following coefficients were obtained: Administrator, .86; Socialite, .83; Teacher, .65; and Researcher, .82. These were regarded as indexes of internal consistency and deemed sufficient for the present study.[1]

RESULTS

Mean scores obtained by subgroups on the Eight College Professors are summarily shown in Table 1. The results of an overall (sexes combined) analysis of variance by Lindquist's (1953, p. 296) Type VI indicated significant professor Type main effects ($p < .001$, $F = 343.35$, $df = 3/891$, $MS_{\text{error}} = 1.18$). Order of preference, as seen in Table 1, was Teacher, Researcher, Socialite, and Administrator. Further, the professor Type interacted significantly with student Level ($p < .05$, $F = 2.54$, $df = 6/891$, $MS_{\text{error}} = 1.18$). Student Level main effects approached the .05 level with greater acceptance shown successively by seniors, graduates, and juniors. In addition, professor Source main effects were significant ($p < .05$, $F = 4.59$, $df = 1/297$, $MS_{\text{error}} = .18$) with education professors receiving higher ratings.

Separate analyses for males and females revealed essentially the same patterns, and no major sex differences among these *S*s were evident in relationship to their faculty preferences. For both sexes the most noticeable difference from the overall results was a nonsignificant professor Source effect.

When the two professor Sources were combined and *S*s' ratings on the four professor Types correlated among themselves, coefficients between the Administrator and the Socialite, Teacher, and Researcher were .25 ($p < .001$), $-.02$, and .14 ($p < .05$), respectively. Those between the Socialite and the Teacher and Researcher were .05 and $-.04$. Finally, the correlation between Teacher and Researcher was $-.15$ ($p < .01$). Since each *S* rated each type, these low correlations are interesting. The direction and size of the coefficients suggest that the image a "teacher" holds in students' minds stands more or less alone apart from any of the other types.

DISCUSSION

In the confines of the present exploratory study, it appeared that students indeed reveal differential preferences among different professor types. In all three groups (male, female, and both), the order of preference was Teacher, Researcher, Socialite, and Administrator. The authors' first prediction was thus partially supported. The second prediction of a larger differential preference among female than male students was not upheld, since both sexes revealed essentially the same preference patterns.

In all three comparisons there were significant ($p < .05$) professor Type × student Level interaction effects. The top rating of Teacher held, but decreased from juniors to graduates.

---

[1] Further analysis by Ebel's method (Guilford, 1954, p. 395) indicated the mean reliability for a single *S* (rater) of .496, while the reliability for the mean of the eight professor ratings from the 300 *S*s was .997.

TABLE 1

MEAN SCORES ON THE EIGHT COLLEGE PROFESSORS BY PROFESSOR TYPE, PROFESSOR SOURCE, SUBJECT LEVEL, AND SUBJECT SEX

Professor		*S*s											
Type	Source	Graduate			Senior			Junior			All levels		
		M ($N = 64$)	F ($N = 23$)	M + F ($N = 87$)	M ($N = 74$)	F ($N = 57$)	M + F ($N = 131$)	M ($N = 37$)	F ($N = 45$)	M + F ($N = 82$)	M ($N = 175$)	F ($N = 125$)	M + F ($N = 300$)
Administrator	A & S[a]	2.89	2.70	2.84	2.59	2.88	2.72	3.16	2.87	3.00	2.82	2.84	2.83
	Educ.	2.89	2.65	2.83	2.59	2.89	2.73	3.22	2.84	3.01	2.83	2.83	2.83
	Both	2.89	2.67	2.83	2.59	2.89	2.72	3.19	2.86	3.01	2.83	2.84	2.83
Socialite	A & S	3.05	3.13	3.07	3.12	3.04	3.08	3.35	3.00	3.16	3.14	3.04	3.10
	Educ.	3.05	3.26	3.10	3.14	3.11	3.12	3.24	3.13	3.18	3.13	3.14	3.13
	Both	3.05	3.20	3.09	3.13	3.07	3.10	3.30	3.07	3.17	3.13	3.09	3.12
Teacher	A & S	4.58	4.57	4.57	4.64	4.49	4.57	4.76	4.78	4.77	4.64	4.61	4.63
	Educ.	4.70	4.48	4.64	4.77	4.56	4.68	4.76	4.80	4.78	4.74	4.63	4.70
	Both	4.64	4.52	4.61	4.70	4.53	4.63	4.76	4.79	4.77	4.70	4.62	4.66
Researcher	A & S	3.30	3.83	3.44	3.04	3.12	3.08	3.05	3.04	3.05	3.14	3.22	3.17
	Educ.	3.28	3.78	3.41	3.16	3.16	3.16	3.03	3.18	3.11	3.18	3.28	3.22
	Both	3.29	3.80	3.43	3.10	3.14	3.12	3.04	3.11	3.08	3.16	3.25	3.20
All types	A & S	3.45	3.55	3.48	3.35	3.38	3.36	3.58	3.42	3.49	3.44	3.43	3.43
	Educ.	3.48	3.54	3.50	3.42	3.43	3.42	3.56	3.49	3.52	3.47	3.47	3.47
	Both	3.47	3.55	3.49	3.38	3.41	3.39	3.57	3.46	3.51	3.45	3.45	3.45

Note.—Scores assigned: strongly reject, 1; reject, 2; neutral, 3; prefer, 4; and strongly prefer, 5.

[a] A & S denotes the College of Arts and Sciences.

Socialite followed suit, but Researcher gained as Level increased. Administrator decreased from junior to senior, but regained his standing with graduates. The authors' third prediction tended to be supported by this evidence. A longitudinal study to examine the suggested sequence is indicated.

The fourth and final prediction of preference for liberal arts over education professors did not receive support. In fact, the opposite preference was shown by these education students.

It was further observed that the ratings were not highly correlated among themselves in spite of their "repeated measurement" feature. A focused nature of student perception is suggested here in agreement with Knapp's (1962) observation discussed earlier.

A limitation of this study is that the results could be a direct function of the specific professor descriptions offered. Nevertheless, the results suggest that, even though the traditional teaching function may have been relegated to secondary status, students still prefer to have a *mentor* as their professor. Administrators may impress them in various other contexts but, so far as classroom instruction is concerned, students prefer a full-time Teacher. A Researcher indeed gains in stature as students climb the educational ladder, but again they would rather study with a Teacher.

The simple instrument, Eight College Professors, seems to be adaptable to many variations in the study of interpersonal perception and preference among educational personnel. Professors' specialty and background experiences may be systematically varied in the descriptions, students' various levels and majors may also be extensively explored, and different schools in different areas may be compared on their students' differential preferences. The instrument is also applicable to groups of faculty members themselves. Some of these possibilities are now being investigated by the authors and will be reported later. Whether the *expressed* reactions correspond with these students' *behavioral* reactions also merits a serious future inquiry.

REFERENCES

CAPLOW, T., & MCGEE, R. The American academic marketplace. In A. H. Halsey, J. Floud, & C. A. Anderson (Eds.), *Education, economy, and society*. New York: Free Press of Glencoe, 1961. Pp. 589–601.

GAGE, N. L. Explorations in the understanding of others. *Educational and Psychological Measurement*, 1953, **13**, 14–26.

GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.

JACOB, P. E. *Changing values in college*. New York: Harper & Row, 1957.

KATZ, J. Personality and interpersonal relations in the college classroom. In N. Sanford (Ed.), *The American college*. New York: Wiley, 1962. Pp. 365–395.

KERR, C. *The uses of the university*. Cambridge: Harvard University Press, 1963.

KNAPP, R. H. Changing functions of the college professor. In N. Sanford (Ed.), *The American college*. New York: Wiley, 1962. Pp. 290–311.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

MCGEE, R. *Social disorganization in America*. San Francisco: Chandler, 1962.

PHILLIPS, D. L. Rejection: A possible consequence of seeking help for mental disorders. *American Sociological Review*, 1963, **28**, 963–972.

PHILLIPS, D. L. Rejection of the mentally ill: The influence of behavior and sex. *American Sociological Review*, 1964, **29**, 679–687.

PUSEY, N. M. The American university 1960. In N. M. Pusey (Ed.), *The age of the scholar*. New York: Harper & Row, 1963. Pp. 155–165.

TANNENBAUM, A. J. *Adolescent attitudes toward academic brilliance*. New York: Teachers College, Columbia University, Bureau of Publications, 1962.

(Received June 17, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 3, 151–158

EFFECTS OF WRITTEN VERBALIZATION AND TIMING OF INFORMATION ON PROBLEM SOLVING IN PROGRAMED LEARNING[1]

ROBERT J. SEIDEL AND IRIS C. ROTBERG
Human Resources Research Office, George Washington University

Training on computer-program (CP) writing was given to 60 high school students in a 3 × 2 factorial design concerned with effects of (a) writing explicitly the rules used in constructing the CPs, (b) writing the names of these rules in conjunction with writing CPs, or (c) writing only the CPs. The other factor was prompting vs. confirmation. Results indicated that: (a) during learning, prompting was significantly superior to confirmation, but a reverse tendency appeared in the criterion tests; (b) naming the rules in addition to writing CPs during training aided later performance when writing more complex CPs on the criterion tests; (c) writing rules during training actually hindered Ss in writing CPs later on the criterion tests.

Much of the research concerning the effects of learning factors in programed instruction (PI) has been conducted with paired associates or serial learning. Little has been done where experimental treatments have been manipulated for hierarchically organized conceptual discourse.[2] Yet this type of discourse is precisely the medium for most, if not all, programed instructional course work.

In a series of serial learning studies by Cook and associates (e.g., Cook & Spitzer, 1960), Angell and Lumsdaine (1961), and others as well, it has been found that prompting is significantly more effective than confirmation for learning. One study by Gagné and Brown (1961) dealing with conceptual material compared several combinations of prompting and confirmation (although they did not use these terms) in the learning of principles for generating number series. In their study, the guided-discovery (confirmation) group, medium-difficulty condition, appeared to be most effective, while the rule-and-example (prompting) condition, low-difficulty level, was least effective.

Thus, Angell and Lumsdaine's (1961) results suggest that the subjects (*S*s) should be given substantial information before being required to respond, while Gagné and Brown's (1961) results suggest that *S*s should be required to discover the answers themselves. The precise factors that caused these experimental differences cannot be determined because of the differences in the paradigms used and because of many other factors that were not comparable across the experiments or between conditions within an experiment. For example, the contiguity of correct stimulus-response association was not held constant across prompting and confirmation conditions, nor were correct responses necessarily given the same number of times nor in the same manner.

---

[1] The research reported in this study was performed by the Human Resources Research Office of George Washington University under contract to the Department of the Army. The contents of this paper do not necessarily represent the official opinion of the Department of the Army. The authors are indebted to Eugene A. Cogan for his helpful suggestions during the preparation of this manuscript. A preliminary version of this paper was read at the American Psychological Association, Los Angeles, September 1964.

[2] A recent study by Hickey and Newton (1964) is a welcomed attempt. Also, for an excellent discussion of problems in doing PI research on conceptual material, see Traub (1964).

For a meaningful comparison of prompting and confirmation, the sole element manipulated should be the amount of information given to *S* prior to requiring him to respond; whatever is not given prior to response should be given after response to make total information given comparable for the confirmation and prompting conditions. A second factor of interest in training for conceptual problem-solving proficiency is the effect of the act of verbalizing, during the training conditions, the rules that are being used in the problem solution.

Paired-associate experiments generally require *S* to make a series of rote responses, and his learning is measured by the ability to give back these rote responses. Conceptual learning, however, requires the abstraction of mediating responses; the success of learning is measured by the ability to apply these mediating skills to a variety of different situations.

Many current PI programs are parallel to paired-associate learning situations in that they seem to emphasize reproducing material presented as written constructed responses. Yet the program developers seem to be interested in teaching concepts (mediating skills). The question that arises then is: Does verbalizing the content of a rule, where separably measurable from mediating responses, aid or hinder a concept-learning and problem-solving capability? Recently, a study by Gagné and Smith (1962) suggested that the verbalization of principles takes on importance as the complexity of the problem increases.

The present study, then, examined the influence of both of these factors—prompting versus confirmation, and alternate ways of verbalizing rules—on learning in a conceptual problem-solving context.

METHOD

Six independent groups of *S*s in a PI environment learned to write increasingly complex computer programs (CPs). To explore the verbalization factor: (*a*) One-third of the *S*s periodically wrote out the content of the rules used to guide the writing of CPs (Rules Group); (*b*) one-third periodically wrote down the names of the rules (Naming Group); and (*c*) one-third wrote CPs without any verbalization of the rules (CP-Only Group). In addition during learning: (*a*) One-half of the *S*s were required to write their answers *after* being given the explicit information needed in the response (prompting)[3]; and (*b*) one-half were required to write their answers *prior to* receiving the information (confirmation). Error rates were obtained during learning on CP writing for all groups and on rules and naming for the appropriate groups.

Treatments were administered during the PI training, and their effects were noted on criterion tests given at the end of each of two phases of the instruction.

Subjects

*S*s consisted of 60 paid volunteer high school students,[4] who had been informed that they were to be paid between $1.50 and $2.50 an hour, depending upon the proficiency they exhibited in learning. *S*s were assigned at random to one of the six conditions ($N = 10$) in a 2×3 factorial design —Prompting versus Confirmation as one factor and the three verbalization conditions as the other. A combination of verbal and clerical elements of the Army Classification Battery was used to classify *Ss* by intelligence level for the detailed evaluation of results.

---

[3] In general, as each new concept was introduced *S* was given a body of information covering its functions. Obviously, then, some degree of global prompting existed for all treatments, so that in the strict sense any prompting-confirmation comparisons are *relative*. However, unless the experimenter waits for his human *S*s to divine the proper "verbal bar press" from an enormous response repertoire, all such studies must of necessity involve a similar event.

[4] The authors are grateful to the principal and staff of Wakefield High School, Arlington County, Virginia, for providing *S*s for this study.

Materials

The material constituted a subcourse of the Programing Specialist Course at the United States Army Signal School (Fort Monmouth, N.J.) and covered primary programing commands of a business-type digital computer. The Rules and Naming Groups used programed instructional booklets consisting of 90 pages, including the criterion tests; booklets used by the CP-Only Group, lacking the rule-verbalization requirement, had half that number of pages. An introduction to symbolic language for computer programing was included in the course. The content was derived from a task analysis of computer programing, with the material hierarchically organized from simple to complex.

The learning task was writing CPs of increasing complexity, with the additional requirements imposed by the experimental treatments. The treatments were applied as each new concept was introduced along with the attendant requirement of writing a CP.

As nearly as possible, the goals of all *Ss* were made comparable by orienting them with a 14-page introduction on course objectives. This orientation included periodic questions which *S* answered covertly until he was satisfied with his understanding of them (cf. Cronbach, 1965).

The criterion tests for Phases I and II consisted of actual computer-programing problems submitted by a programers' group from Fort Huachuca (Ariz.) and by programing instructors from the Fort Monmouth Signal School. Where necessary, problems were simplified by removing aspects that would require techniques beyond the scope of the course. As a secondary part of the criterion tests, the ability of all *Ss* to write the rules and the names of the rules which they used in writing the CPs was also measured.

The sample rule given below is illustrative of the Rules and Naming treatments:

> The CLA SOCSEC instruction clears any material already in the accumulator and copies the number from location SOCSEC into it.

The portion underlined above was given to the Rules Group, and these *Ss* were to fill in the rest, whereas the nonunderlined portion was given to *Ss* in the Naming Group, and they filled in the part underlined. Since *S* worked on only one page at a time, the Prompting-Confirmation variable can also be illustrated. For the Prompting Group the entire sentence was presented first; *S* turned the page and completed the blank. The opposite held for the Confirmation Group.

Procedure

The course was administered in two phases. On Day 1, students were given Phase I, the section concerned with primary programing commands and a criterion test for that phase. On Day 2, they were first given a retention test consisting of four problems to be programed; this was not a timed test, but all *Ss* were through within 30 minutes. Immediately following retention testing, they were given Phase II of the course, which covered more complex CPs, and the criterion test for that phase. On Day 3, Retention Test 2, four programing problems covering all the course material, was given, followed by the intelligence measure.

The students were allowed a maximum of 4 hours on each of the 2 days of training, but only one or two students required that much time on either day. Booklets were checked for each student at the end of his working day to ensure that he was actually working with the material and not simply staying for his pay. All students appeared highly motivated to perform; many stayed after hours to ask further questions about the course work. (No substantive questions were answered, to avoid contaminating experimental treatments.)

RESULTS

Evidence for Learning

On the test for Phase I, nearly two-thirds of the *Ss* achieved 80% correct or more on the program-writing material, and more than a third of the *Ss* scored 90% or more. Results for Test 2 were comparable. These scores on the criterion tests for the principal dependent variable—computer programing—and scores on the retention tests are given in Table 1. Also included are results of Bartlett tests on the variances.

TABLE 1

CRITERION AND RETENTION TEST MEANS AND VARIANCES BY TREATMENTS

Test	Training group			Summary mean	% correct	*SD*		
	Rules + CP	Naming + CP	CP -Only			Rules group	Naming group	CP-Only group
Criterion Test 1[a]								
Prompting	39	43	47	43	83	14.84	8.66	7.35[b]
Confirmation	43	49	44	45	87	10.25	3.61	7.55
Summary mean	41	46	46					
% correct	79	88	88					
Criterion Test 2[c]								
Prompting	66	66	75	69	78	28.76	15.75	10.63[d]
Confirmation	69	79	77	75	84	19.44	7.62	11.96
Summary mean	68	73	76					
% correct	76	82	85					
Retention Test 1 (%)	76	87	84		82		*ns*	
Retention Test 2 (%)	70	80	68		73		*ns*	

[a] Maximum possible score: 52.

[b] $\chi^2(5) = 15.52, p < .01$.

[c] Maximum possible score: 89.

[d] $\chi^2(5) = 18.16, p < .01$.

In all cases, intelligence, as indicated by scores on the Army Classification Battery, was significantly and positively correlated with the criterion scores. No interactions appeared between intelligence level and the experimental treatments on the principal measure, CP writing.

Treatment Effects

Learning time. For both Phases I and II, prompting and confirmation treatments did not differ significantly in *time* to complete the course. The verbalization groups did differ significantly on Phase I ($F = 5.391$) and on Phase II ($F = 3.798$). Table 2 gives the mean time scores for the verbalization-treatment conditions. The treatment variances did not differ significantly.

TABLE 2

LEARNING-TIME MEANS BY VERBALIZATION TREATMENTS

	Means (in min.)		
	Rules + CP	Naming + CP	CP-Only
Phase I[a]	134.5	116.3	115.1
Phase II[b]	107.0	92.6	101.6

[a] $F(2.54) = 5.391, p < .01$.

[b] $F = 3.798, p < .05$.

Significance of the F ratios in Phase I for the verbalization variable was due principally to the fact that the Rules Group took longer to complete the material than did the other two verbalization groups. In Phase II the Naming Group was faster than the other two groups.

Since the CP-Only Group did not have the rules or naming items in their programed instructional materials, they had about half as many booklet pages as the other two groups. Despite this, the CP-Only Group did not finish faster than the Naming Group in either Phase I or Phase II. Because of this fact and because achievement generally was so high

TABLE 3

TIME-WEIGHTED MEANS ON CRITERION: PROMPTING VERSUS CONFIRMATION

Measure	Criterion Test 1		Criterion Test 2		Retention Test 1		Retention Test 2	
	Prompting	Confirmation	Prompting	Confirmation	Prompting	Confirmation	Prompting	Confirmation
CP writing	368	389	719	750	427	442	423	468
Naming the rules	34	38	49	46				
Stating the rules	17	24	21	19				

Note.—Score = $\frac{\text{Achievement}}{\text{Time}} \times 1{,}000$.

and positively correlated with speed of learning, further analyses of variance on the criterion tests were done with a time-weighted score for each individual, computed by dividing *S*'s criterion score by his time to complete the course.

Criterion scores. The general treatment effects on the time-weighted scores were as follows:

1. Prompting-Confirmation Results. During learning, the prompting condition was significantly superior to confirmation. However, on the criterion tests the groups did not differ significantly, and, if anything, the Confirmation Group tended to be superior, as shown in Table 3. The same tendency holds for the retention tests.

Comparing prompting versus confirmation across the three intelligence levels and counting (on all criterial scores, including Rules and Names) the number of times the Confirmation Group's means exceeded those of the Prompting Group (using Test 1, Test 2, and both retention tests in the count), a ratio of 17 to 6 superiority of the Confirmation Group was found. During learning, on the other hand, the error rates were: Phase I, program writing—Prompting = 3%, Confirmation = 10%; rules and naming—Prompting = 6%, Confirmation = 24%. In Phase II, program writing—Prompting = 2%, Confirmation = 5%; rules and naming—Prompting = 10%, Confirmation = 44%.

2. Verbalization Results. Turning to the *F* ratios for the other variable, verbalization, it is found that *S*s required to write the content of the rules during training did not do as well in writing CPs on the criterion tests as *S*s who had simply to write the names of the rules during training or *S*s who learned without either additional requirement. It should be noted that, as shown by the unweighted means in Table 1, the Rules Group also had the highest variance on the criterion tests. (This effect was particularly evident with *S*s of lower intelligence.)

Table 4 shows that if we order the means of the three groups on CP writing—the criterion measure of principal interest—the mean of the Rules Group is consistently lower than the mean of the Naming Group or the CP-Only Group. Two-group comparisons (Scheffé, 1959) indicate this difference is the basis for statistical significance in all cases.

An area for further study is suggested by the change in the ordering on the program-writing measure between Test 1 and Test 2 for the Naming Group versus the CP-Only Group. On Test 1 and Retention Test 1

TABLE 4

TIME-WEIGHTED MEANS ON CRITERION: VERBALIZATION VARIABLE

Measure	Test 1			Test 2			Retention Test 1			Retention Test 2		
	CP-Only	CP + Rules	CP + Naming	CP-Only	CP + Rules	CP + Naming	CP-Only	CP + Rules	CP + Naming	CP-Only	CP + Rules	CP + Naming
CP writing	426	314	397	753	624	793	486	356	463	412	385	521
Naming the rules	40	29	40	42	39	55						
Stating the rules	20	24	20	15	27	19						

Note.—Score = $\frac{\text{Achievement}}{\text{Time}} \times 1{,}000$. All three-group comparisons except Stating the rules, Test 1, were significant ($p < .05$). For Naming the rules, Test 2, $p < .01$.

the CP-Only Group made the highest scores, while on Test 2 and Retention Test 2 the Naming Group seemed to be superior. This tendency, however, is *not* statistically significant. This finding is being explored in a larger study dealing with the same introductory content, but with more complex concepts of computer programing as well. It may be, as noted in a recent study by Gagné and Smith (1962), that verbalization of principles is not important except for more complex material.

We should also note that the inferior performance by the Rules Group in CP writing was not simply a result of poorer learning *generally*. The Rules Group was superior to the other groups on the criterion requiring the writing of rules. In Phase I this was not significant, but in Phase II it was ($p < .05$). Thus the Rules Group learned the verbal materials better, but at the expense of not having learned how to use them. The Naming Group, on the other hand, answered the naming questions on the criterion significantly better ($p < .05$) than the other two groups and, consistent with the Gagné and Smith finding, benefited from this added background in CP writing on Phase II.

One final finding was that the criterion measures were highly reliable by Hoyt's test (cf. Johnson, 1949). On CP writing the coefficients ranged from 79 to 89 ($p < .01$).

DISCUSSION

The data from the prompting versus confirmation comparison indicate the necessity for caution in applying findings established in paired-associate or serial learning to paradigms involving a highly organized, abstracted learning, such as learning how to write CPs. In the latter case, the material being taught is a mediating set of knowledges rather than mechanical rote responses which *S* performs during the course itself (as in paired-associate learning). In the present study it was necessary for *S* to extract from his mechanical writing responses the appropriate mediating stimulus-response connections required to handle the conceptual problem of writing.

Viewed in another light, the prompting-confirmation data also carry implications for the relation between error rate during learning and success on the criterion. The present data indicate that *in the present context* the effect of error rate during learning washed out or even tended to be negatively related to errors in end-of-course testing. Other suggestive evidence for this relationship is found in a recent evaluation of a PI course for training salesmen (Basic Systems Incorporated, 1963). The authors' calculation of the relationship of error rate during learning and percentage of errors on the criterion test showed a Spearman ρ of $-.32$, which was not quite significant at the .05 level (but would have been for a one-tailed test), but certainly agreed with the current finding. One possible reason for not getting a larger ρ was the combination of a small sample ($N = 27$) and a ceiling effect in the criterion scores, which ranged between 90–100% correct in all but two cases.

Secondly, with respect to verbalization, what is in a name—or in a rule—is definitely important. The *ways* in which we ask *S* to verbalize (e.g., either by writing, as was the case in this study, or by oral recitation) must be taken into account when considering the effects of this technique. Apparently, requiring the use of written mnemonics (the names of rules) during learning enabled the Naming Group to learn the rules as well as how to use them. On the other hand, given the contextual restraint as the Rules Group was, and requiring storage of the entire rule, prevented *Ss* from also learning efficiently how to use the rules in solving computer problem.

Previously, Gagné and Smith (1962) found an advantage in requiring *S* to compose and verbalize the rules he was using in a concept-identification task. However, in the present case the rules *S* was required to write differed in that they were contextually constrained. First, *Ss* were required to paraphrase the rule closely in a completion format; second, if a computer command were involved, *S* was not simply asked the general principle behind the command, but rather he was asked to interpret a computer instruction with the command applied to a specific meaningful program context.

Requiring *Ss* to write contextual rules in this manner, in addition to solving problems, should indicate the value of teaching *Ss* to reiterate the concepts mediating the specific problem-solution performance which they are being taught. The naming responses that were required in the problem-solving performances should indicate the value of the use of mnemonics or reduced cues as an aid to learning the performance-oriented response (computer programing in this study).

In attempting to account for the possible advantage of having *S* name the rule as opposed to repeating a stereotyped set of words comprising a rule as the experimenter has given it to him, the possibility must be taken into account that *S* will perform best when he can use his own repertoire of verbal responses. Gagné and Smith provided the widest possible latitude in this framework by allowing *S* to compose the rule himself as he solved a series of problems, and this procedure seemed to help. Ericksen (1962) has also advocated this approach. Clearly, an evaluation of the importance of this problem for conceptual learning merits further investigation.

REFERENCES

ANGELL, D., & LUMSDAINE, A. A. The effects of prompting trials and partial-correction procedures on learning by anticipation. (Research Report AIR-C14-9/61-SR6) Pittsburgh, Pa.: American Institute for Research, September 1961.

BASIC SYSTEMS, INCORPORATED. *WATS: Development and field testing: Report of* *the development and installation test data from a programmed self-instructional training course.* (Technical Report developed for AT&T) New York: Basic Systems Incorporated, 1963.

COOK, J. O., & SPITZER, M. E. Prompting versus confirmation in paired-associate learning. *Journal of Experimental Psychology,* 1960, **59,** 275–276.

CRONBACH, L. J. The logic of experiments on discovery. Paper read at Social Science Research Council, Committee on Learning and the Educational Process Conference, New York, January 1965. (Stanford University, United States Office of Education, Cooperative Research Program).

ERICKSEN, S. C. The place of thinking in an ideal university. *American Psychologist,* 1962, **17,** 763–771.

GAGNÉ, R. M., & BROWN, L. T. Some factors in the programing of conceptual learning. *Journal of Experimental Psychology,* 1961, **62,** 313–321.

GAGNÉ, R. M., & SMITH, E. C., JR. A study of the effects of verbalization on problem solving. *Journal of Experimental Psychology,* 1962, **63,** 12–18.

JOHNSON, P. O. *Statistical methods in research.* New York: Prentice-Hall, 1949.

HICKEY, A. E., & NEWTON, J. M. The logical basis of teaching: I. The effect of subconcept on learning. Office of Naval Research Report, January 1964, Contract Nonr-42 15(00).

SCHEFFÉ, H. *The analysis of variance.* New York: Wiley, 1959.

TRAUB, R. E. *The importance of problem heterogeneity to programed learning.* (Research Bulletin 64-26) Princeton, N. J.: Educational Testing Service, 1964.

(Received June 28, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 3, 159-163

EFFECT OF EXPERIMENTALLY LEARNED WORD ASSOCIATIONS ON THE ACQUISITION OF READING RESPONSES[1]

S. JAY SAMUELS
University of Minnesota

Pairs of words were learned by 44 1st graders during word-association training. They were then given reading training on the same or different pairs of words and tested for word recognition and speed of recognition on the 2nd word of each word pair. In the facilitation treatment, where word pairs were the same for word-association and reading training, the mean recognition score was 1.46. In the interference treatment and neutral treatment, where word pairs were different for word-association and reading training, the means were .84 and .82. In the control treatment, where no word-association training was given, the mean was .34. These differences were significant ($p < .01$), and higher word-recognition scores were associated with faster speeds of recognition ($p < .01$).

Reading specialists and psychologists have stated that performance on verbal tasks is improved with either stimulus or response familiarization. Muehl (1960) has emphasized the importance of stimulus familiarization by demonstrating positive transfer in learning to read a list of words when visual discrimination pretraining was given with the same words used later in reading training. Strickland (1962), Ruddell (1965), and Lefevre (1964) have also stressed the need for familiarization in reading by stating that primers should be written using the same syntactical patterns children use in speaking. Having children dictate their own reading material, as in the experience-chart method, is another means of achieving syntactical familiarization (Gans, 1963; Goodman, 1963). Dolch (1960) believes that for many beginning readers, the course of learning to read from experience charts consists of first memorizing the sequence of verbal responses which appear on the chart, followed later by learning to read the words, that is, by associating the verbal response with the printed word stimulus.

The facilitating effect of familiarization has also been demonstrated with verbal tasks in which the subject (*S*) is required to recognize flashed words. As the amount of visual preexposure to nonsense words increases, the time necessary to recognize these words decreases when they are presented with a tachistoscope (Solomon & Postman, 1952; Sprague, 1959). Miller, Bruner, and Postman (1954) found that when nonsense words are constructed with different orders of approximation to English, the higher the order of approximation, the lower the visual recognition threshold, and Tulving and Gold (1963) note that as the amount of information in a sentence containing a missing word increases, the time required to recognize the missing word decreases.

The positive effect of familiarization has also been demonstrated with word associations. O'Neil (1953) and Rouse and Vernis (1963) demon-

---

[1] This study is based upon a dissertation submitted to the University of California at Los Angeles in partial fulfillment of the requirements for the EdD degree. The writer would like to express his appreciation to his chairman, Merlin C. Wittrock, and to Norman Anderson, Phillip Gough, and David Wiley for their helpful advice and assistance.

TABLE 1
PARADIGM TO ILLUSTRATE EXPERIMENTAL DESIGN

Word-association familiarization	Reading training		Word-recognition test order
S_1–R_1 BLUE-dominoes	S_1–R_1 BLUE-dominoes	F	R_3 thumbnail
S_2–R_2 PLAY-fairy	S_2–R_2 PLAY-faerie		R_2 faerie
S_3–R_3 EAT-thumbnail	S_3–R_4 EAT-bumblebee	I	R_8 chequer
S_4–R_4 BIG-bumblebee	S_4–R_3 BIG-thumbnail		R_4 bumblebee
S_9–R_5 BLACK-music	S_5–R_5 RED-muzjik	N	R_6 liquorice
S_{10}–R_6 BUY-liquorice	S_6–R_6 MAKE-liquorice		R_1 dominoes
None	S_7–R_7 BLACK-martial	C	R_7 martial
None	S_8–R_8 LIKE-chequer[a]		R_5 muzjik

[a] Pronounced "checker."

strated that when word associates are tachistoscopically exposed in succession, recognizing the first word aids in recognizing the second word, and the stronger the association between the words, the lower the recognition threshold.

The purpose of the present study was to investigate what effect experimentally learned word associations would have on learning to read the second word in a two-word chain when only minimal help was given during reading training. All *Ss* were able to read the first word in the chain, whereas none were able to read the second word at the start of the experiment.

METHOD

Ss in the investigation were 48 first graders from a Los Angeles elementary school who were randomly selected and assigned to one of four levels of the design.

A 4 × 4 Latin-square design was used. Each *S* received four treatments in succession. For each level of the design, the order of treatments and the word associates were counterbalanced.

As seen in Table 1, in the facilitation (F) treatment, the word associates learned during familiarization were S_1–R_1 and S_2–R_2. During reading training, S_1–R_1 and S_2–R_2 were presented.

In the interference (I) treatment, S_3–R_3 and S_4–R_4 were learned. During reading training, S_3–R_4 and S_4–R_3 were presented.

In the neutral (N) treatment, S_9–R_5 and S_{10}–R_6 were learned. During reading training, S_5–R_5 and S_6–R_6 were presented.

In the control (C) treatment, no word-association familiarization was given. During reading training, S_7–R_7 and S_8–R_8 were presented.

Ss worked individually with the experimenter (*E*). They were first pretested on all the stimulus and response words used in reading training, then given word-association familiarization on six pairs of words, then reading training on eight pairs of words, and a final test of word recognition.

Pretest

To test the effect of word associations on learning to read the second word in each word pair, it was essential that *Ss* be able to read the first (stimulus) word but not the second (response) word at the start of the experiment. Therefore, each *S* was pretested on eight stimulus and eight response words used in reading training (see Table 1). The stimulus words were at the preprimer level, whereas the response words were from the Thorndike-Lorge (1944) list of words which appear less than once in a million times in print.

All visual stimuli used in this experiment were typed with a primary-style typewriter on 5 × 8-inch index cards.

The pretest consisted of presenting the cards one at a time to *S* for 20 seconds and asking him to read the word typed on the card. If he failed to read any of the stimulus words, or if he read any of the response words, he was eliminated.

Four *Ss* were eliminated, one from each level of the design.

Word-Association Familiarization

After the pretest, word-association familiarization was given. As seen in Table 1, *S* learned six pairs of words in succession. For the first pair, *E* and *S* alternately said the word pair five times. For example, *E* said, "BLUE-dominoes," then *S* said, "BLUE-dominoes." Then *E* alternately gave the stimulus word and *S* gave the response word five times. For example, *E* said, "BLUE," and *S* said, "dominoes." Following this, the next pair of words was learned in a similar manner. Then a review of the first pair was begun. During the review, *E* alternately said the stimulus word and *S* gave the response word five times. After the review of the first word pair, the third pair of words was learned. The process was continued for each of the six pairs of words.

Reading Training

Following familiarization, reading training was given in two stages. In Stage 1, each pair of words was presented once, on a card similar to Figure 1A. At each presentation *E* read the pair of words at the top of the card. *S* then pointed to the word pairs at the top and read them. Following this, *S* pointed to the multiple-choice word pairs at the bottom which matched the word pairs at the top and read them.

After *S* read the eight pairs of words with help from *E*, Stage 2 was begun. During this stage, at no time did *E* read any words for *S*. As seen in Figure 1, Parts A, B, and C, three different kinds of reading-training cards were presented in sequence. *S* was instructed to point to the words at the top of the card and read them, then to select the matching words at the bottom and read them.

For each word pair, six cards were used so that if *S* was able to read the words, he went from emitting oral responses for the stimulus and response words to emitting an oral response for the response word alone.

For the C treatment, seven additional cards like Figure 1A were used to provide the same number of opportunities to respond during reading training that were missed during word-association familiarization.

Word-Recognition Test

Immediately following reading training, a word-recognition test was given on the eight response words. Two measures of learning were used—number of correct reading responses, and speed of recognition. The test for each treatment was 40 seconds long. Each test consisted of two words, presented one at a time for 20 seconds.

RESULTS

Number of Correct Reading Responses

As seen in Table 2, the mean number of correct reading responses given on the word-recognition test was .34 for the C treatment, .82 for the N treatment, .84 for the I treatment, and 1.46 for the F treatment.

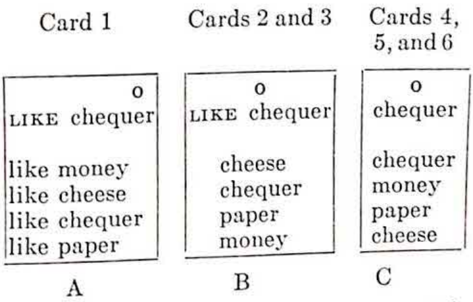


FIG. 1. Examples of three kinds of cards used in reading training.

The analysis of variance for repeated measure Latin squares (Winer, 1962) disclosed that neither the order of treatment presentation nor the interaction between treatments and list was significant. The main treatment effect of word-association familiarization on learning reading responses was significant ($F = 25.47$, $df = 3/120$, $p < .01$).

The Newman-Keuls tests on pairs of means (Winer, 1962) indicated that on the word-recognition test the F treatment was superior to the I treatment ($p < .01$), the N treatment ment ($p < .01$), and the C treatment ($p < .01$). The N and I treatments were both superior to the C treatment ($p < .01$). The I treatment was not significantly different from the N treatment.

Speed of Recognition

As seen in Table 3, the mean recognition speed in seconds for the two words on each test was 34.66 for the C treatment, 28.29 for the N treatment, 27.50 for the I treatment, and 17.93 for the F treatment.

The analysis of variance for repeated measure Latin squares (Winer, 1962) disclosed that neither the order of treatment presentation nor the word-association list used nor the interaction between treatments and list

TABLE 2

MEAN NUMBER OF CORRECT READING RESPONSES AND STANDARD DEVIATIONS ON THE WORD-RECOGNITION TEST

	Treatment			
	Control	Neutral	Interference	Facilitation
M	.34	.82	.84	1.46
SD	.30	.68	.80	.69

was significant. The main treatment effect of word-association familiarization on speed of recognition was significant ($F = 23.01$, $df = 3/120$, $p < .01$). The Newman-Keuls test on pairs of means (Winer, 1962) indicated that speed of recognition was faster for the F treatment than the I treatment ($p < .01$), the N treatment ($p < .01$). The recognition speed was faster for the N treatment than the C treatment $p < .05$), and the I treatment was faster than the C treatment ($p < .01$). The speeds of recognition for the I and N treatments were not significantly different from each other.

DISCUSSION

The results of this experiment indicate that experimentally learned word associations can affect the acquisition of reading responses. The highest word-recognition score and fastest speed-of-recognition score was found in the F treatment in which the same word pairs were presented for word-association familiarization and reading training. A second finding was that even in the I and N treatments, where the word associations conditioned during familiarization were not designed to facilitate reading acquisition, superior learning occurred when compared to the C treatment, in which no word-association familiarization was given.

TABLE 3

MEANS AND STANDARD DEVIATIONS IN SECONDS FOR SPEED OF RECOGNITION

	Treatment[a]			
	Control	Neutral	Interference	Facilitation
M	34.66	28.29	27.50	17.93
SD	9.23	10.67	12.46	13.35

[a] Each word was exposed for 20 seconds on the two-word test.

The familiarization effect may be used to explain why there were superior test-performance results with all treatments in which word associations were learned. The response word, which was the second word in each of the six word pairs memorized during familiarization, was presented during reading training and on the word-recognition tests. These learned response words were available to *S* either for learning during reading training or guessing on the word-recognition test. In the C treatment, where no word-association familiarization was given, although *S* was told the response word once, there was sufficient forgetting, so that in most cases he was unable to recall the response word and consequently was unable to learn to read the word used in that treatment.

On the word-recognition test, a score for a correct reading response might have come about by *S* reading the word because he had learned it or by guessing. Each *S* had 20 seconds for each word on the test. If, on the test, *S* tried to "guess" the correct response by going through the repertoire of response words conditioned during familiarization, the probability of his emitting the correct response was .75. A few *S*s gave evidence of following this guessing strategy. This strategy is often used by beginning readers when unable to read a word such as "ate" in a sentence such as: "The boy ate his food." Because of prior learning of syntactical habits, the student emits verbs. He may try a number of them until one of them fits the context of the sentence. Although this appears to be a guessing procedure, it is not random guessing but cued guessing. The function of contextual cues and grammatical habits is to limit verbal responses to those which have a high probability of being correct. Similarly, in this experiment, one of the functions of familiarization was to provide *S* cues which increased the probability of his guessing the correct response when additional information was needed.

The speed-of-recognition scores were used in this experiment to determine the effect of guessing. Guessing could be expected to increase the time required for word recognition. Since the treatments with the highest scores for word recognition were associated with the fastest speeds of recognition, one must conclude that although *S*s could guess on the test, high word-recognition scores were not primarily the result of guessing.

REFERENCES

DOLCH, E. W. *Teaching primary reading.* Champaign, Ill.: Garrard Press, 1960.

GANS, R. *Common sense in teaching reading.* Indianapolis, Ind.: Bobbs-Merrill, 1963.

GOODMAN, K. S. A linguistic study of cues and miscues in reading. *Elementary English*, 1965, **42,** 639–643.

LEFEVRE, C. A. *Linguistics and the teaching of reading.* New York: McGraw-Hill, 1964.

MILLER, G. A., BRUNER, J. S., & POSTMAN, L. Familiarity of letter sequences and tachistoscopic identification. *Journal of General Psychology,* 1954, **50,** 129–139.

MUEHL, S. The effects of visual discrimination pretraining on learning to read a vocabulary list in kindergarten children. *Journal of Educational Psychology,* 1960, **51,** 217–221.

O'NEIL, W. M. The effect of verbal association on tachistoscopic recognition. *Australian Journal of Psychology,* 1953, **49,** 333–338.

ROUSE, R. O., & VERNIS, S. J. The effect of associative connections on the recognition of flashed words. *Journal of Verbal Learning and Verbal Behavior,* 1963, **1,** 300–303.

RUDDELL, R. B. The effect of oral and written patterns of language structure on reading comprehension. *The Reading Teacher,* 1965, **1,** 270–275.

SOLOMON, R. L., & POSTMAN, L. Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology,* 1952, **43,** 195–206.

SPRAGUE, R. L. Effects of differential training on tachistoscopic recognition thresholds. *Journal of Experimental Psychology,* 1959, **58,** 227–231.

STRICKLAND, R. G. The language of elementary school children: Its relationship to the language of reading textbooks and the quality of reading of selected children. *Bulletin of the School of Education.* Bloomington: Indiana University, July 1962.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words.* New York: Teachers College, Columbia University, Bureau of Publications, 1944.

TULVING, E., & GOLD, C. Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology,* 1963, **66,** 319–327.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received July 7, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 3, 164–168

CONCEPT DEVELOPMENT IN PIAGETIAN TERMS IN RELATION TO SCHOOL ATTAINMENT[1]

P. S. FREYBERG
Massey University

Relationships between general intellectual ability, conceptual development and attainment in arithmetical computation, and arithmetical problem solving and spelling were investigated in a 2-yr. longitudinal study involving 151 New Zealand children aged 6–9 yr. Conceptual development, measured by a 72-item objective test, was more highly correlated with Primary Mental Abilities Test mental age ($r = .523$) than with chronological age ($r = .116$). Regression analysis showed that the inclusion of concept-development scores with mental age accounted for a significantly larger proportion of attainment score variance in most cases. It is suggested that children's school performance is associated with aspects of conceptual development not adequately assessed by conventional intelligence tests.

Although attention has recently been directed (Braun, 1963) toward the relationship between children's ability to form concepts and their school attainment, comparatively little research has been reported which relates school attainment to developmental levels in conceptual thinking. The efforts of Piaget and his Geneva colleagues have been concerned mainly with the ways in which conceptual thinking develops, rather than with the educational implications of ontogenetic changes in conceptualization.

Piaget's hypothesis of discontinuous stages in concept development is now comparatively well known, if not widely accepted, in English-speaking countries. Detailed expositions have been given by Hunt (1961) and Flavell (1963), and in translations of Piaget's (particularly 1950, 1952a, 1952b) own works. It is sufficient to refer here to Piaget's contention that profound changes in children's capacities for conceptual thinking occur at about the ages of 7–8 and 11–12 years. He maintains that before the age of 7 years children cannot think "reversibly" and so are not able to form adequate concepts of classes, relationships, and numbers, and that, until the age of 11 years or so, the concepts which they do form are so related to concrete situations that they cannot handle abstract problems. Although he recognizes the influence of environmental factors on the ontogenesis of conceptual thinking, his main emphasis has been on maturational determinants.

Some other research workers (e.g., Bruner, 1960; Churchill, 1958a, 1958b) have suggested, however, that it may be possible through the use of appropriate educational methods to speed up the rate of concept development, and that this might then allow the earlier teaching of important educational skills and knowledge. The assumption is made that some skills taught in school—particularly those requiring abstract verbal thinking—cannot be attained until the appropriate level of conceptualization has also been achieved. While this assumption is reasonable, the extent of the dependence of school attainment on prior concept development has not yet been adequately explored.

---

[1] This study is part of a larger longitudinal investigation of cognitive development in children aged 6–9 years, supported by a research grant from the New Zealand University Grants Committee. The assistance of A. E. Fieldhouse (Victoria University, Wellington) and the staffs of the schools concerned is gratefully acknowledged.

Two experimental studies which indicate the need for further investigation have been those of Dodwell (1961) and Hood (1962). Dodwell constructed a standardized test of number concepts—conservation, correspondence, seriation, cardination, and ordination—along Piagetian lines, but with greater objectivity than is possible using Piaget's "clinical" methods. Short-term stability (test-retest) correlation with an interval of 1 week was reasonably high at $r = .87$. When the scores of 34 kindergarten children on this test were compared with their subsequent performance in a simple test of arithmetical achievement at Grade 1, 6 months later, the correlation was .59.

Hood (1962) devised tests of concept development involving conservation of number and quantity, correspondence, seriation, and additive composition of classes and numbers, again similar to some of Piaget's experiments. He compared the total scores of 126 children aged 4–8 years with teachers' ratings of the arithmetical skills of their pupils and found a substantial degree of correspondence. Most of those who were successful in the Piagetian tests were found to be making reasonable progress in number work, while there was "no instance of success in arithmetic and failure in the Piaget tests being found in the same child [p. 284]."

Unfortunately, the results of both of these investigations can be regarded only as starting points: Dodwell's group of subjects was too small, and Hood relied on teachers' ratings of arithmetical attainment rather than on actual performance. The purpose of the present study, therefore, was to investigate in greater depth the relationship between children's level of concept development and their school attainment in three areas—skills in arithmetical computation, in solving arithmetical problems, and, in the verbal field, in spelling.

METHOD

Subjects

The subjects (*S*s) of the investigation were 151 children (64 boys, 87 girls) from four schools in a relatively prosperous New Zealand inland urban center with a total population of about 42,000 people. Although socioeconomic differences between the areas in which each school is situated are not marked, such differences as did exist were taken into account in order to avoid bias in the sample. As some of the tests involved reading of answer choices (which were, however, always read aloud to the testees), children who were, after a year's schooling, complete nonreaders were excluded from the sample. The sample may thus represent a distribution slightly curtailed at the lower levels of intelligence, although the range of IQs was 77–133, with a mean of 104, on the Primary Mental Abilities (PMA) 5–7 year test (United States norms).

The chronological age (CA) range at initial testing was 5 years, 9 months to 7 years, 10 months, with a median of approximately 6 years, 9 months. An experimental group of 80 children was tested at three-monthly intervals over 2 years, while the control group of 71 children was tested only at the beginning and end of this period.

Tests and Procedures

A 72-item objective test of concept development was first devised and evaluated (Freyberg, 1964). Eighteen categories, with four items to each category, included tests of conservation of quantity and weight, of numerical correspondence, of additive composition of numbers and classes, and of concepts of position in space, speed, age, kinship, and causal relationships. All but five items were of the multiple-choice type, most of the items being illustrated by simple diagrams. The multiple-choice questions required the checking of one of three simple statements, both questions and answer choices being read to *S*s.

For the age range concerned (6–9 years

TABLE 1

INTERCORRELATIONS AMONG VARIABLES AT FIRST TEST PERIOD

Variable	MA	Concept scores	Spelling scores
CA	.276	.116	.127
MA (PMA 5–7)		.523	.462
Concept-development scores			.418

Note.—$N = 151$.

inclusive) the test-retest reliability with 1 week's interval was .911 ($N = 99$), and the internal consistency coefficient (split-half corrected) was .869. The validity of the objective test as a group instrument was assessed by comparing the results from 17 items with the scores obtained on corresponding Piagetian tests in a clinical situation. The correlation was .722 with $N = 101$.

The attainment tests consisted of a speeded computational test of 120 items involving addition, subtraction, multiplication, and division (retest reliability .928, $N = 70$); a 25-item test of arithmetical problem solving (reliability .887, $N = 68$); and a 22-item dictated-word spelling test (reliability .963, $N = 101$). A conventional measure of intelligence, the PMA Test for ages 5–7 and 7–11, was also used.

The concept development, spelling, and intelligence tests were administered at the initial test period to all *S*s, who were then allocated within each school to experimental or control groups so as to secure samples matched for intelligence. Differences between groups on the other variables proved to be statistically insignificant. The tests of computation and arithmetical problems were introduced into the test program midway through the series for the experimental group, but were taken by the control group only once, that is, at the final test period.

At the end of 2 years, owing to transfers away from the schools and absences at the final test administration, the number of *S*s had dropped to 124. Significant practice effects were observed in the experimental group with the tests of spelling, concept development, and intelligence, but not with the arithmetical tests. In examining the predictive relationships between concept development and arithmetic skills it was thus possible to utilize the results of both control and experimental groups, but the concurrent relationships at the final test period had to be determined from the results of the control group alone.

Relationships between the variables were assessed by means of product-moment correlations, the significance of the contribution of concept development being estimated through the calculation of multiple-correlation coefficients and regression analysis of variance.

RESULTS

The intercorrelations among variables at the initial test period are given in Table 1. It will be observed that the correlation of concept scores with mental age (MA) is considerably higher than it is with CA. The CA range was not in fact so restricted that it would have effectively prevented the demonstration of any significant relationships; the youngest child was at that time aged 5 years, 9 months, the oldest 7 years, 10 months, and the quartiles were 6 years, 5 months and 7 years, 1 month. The correlation of concept scores with CA when MA was held constant ($r = .08$) was not significantly different from zero, whereas the correlation between concept development and MA with CA held constant was $r = .52$.

The multiple regression of concept scores and MA together on spelling scores produced a coefficient of $r = .504$. Regression analysis of variance showed that the inclusion of concept scores with MA accounted for 4% of variance over and above the 21% accounted for by MA alone, a small but statistically significant proportion ($F = 8.20$, $df = 1/148$; $p < .01$).

The relative predictive value of the tests of general intellectual ability and concept development as far as later arithmetic attainment 2 years later was concerned is demonstrated in Table 2. The differences between the first-order correlation coefficients are statistically nonsignificant, and the coefficients are only moderately high, as is usual with predictive calcula-

TABLE 2

CORRELATIONS BETWEEN COGNITIVE VARIABLES AT INITIAL TESTING AND ATTAINMENT 2 YEARS LATER

Variable	Arithmetic computation[a]	Arithmetic problems[b]
MA (PMA 5–7)	.543	.557
Concept scores	.524	.571
Combined MA/Concept	.612	.647

[a] $N = 121$.
[b] $N = 124$.

tions at this age level. Nevertheless, regression analysis of variance brought out the fact that, in both cases, consideration of concept scores in addition to MA added significantly to the predictive power of MA alone. The relative variance contributions are shown in Table 3.

The corresponding data relating to the prediction of spelling attainment making use of the results of the control group alone showed a smaller degree of correlation ($r = .397$ for MA, $r = .454$ for concept scores). Multiple regression failed to produce a significant increase in predictive power.

An examination was made of the concurrent relationships between cognitive and attainment variables at the final test period, again excluding results from the experimental group because of practice effects in some tests. The consideration of concept scores accounted for an additional 8% of variance, over and above 52% for MA, in a multiple regression on arithmetical problem performance—significant at the .01 level. Concept scores made no significant contribution to the variance in the case of arithmetical computation or spelling.

DISCUSSION

These results provide confirmation for the findings of such British and North American research workers as Carpenter (1955), Beard (1960a, 1960b), Elkind (1961), and Feigenbaum (1963) that concept development is more closely linked to the growth of general intellectual ability than just to CA (or general maturational level). It is interesting to observe that the reported correlations of concept development with MA have in general been higher than those with IQ, as one would expect if CA in fact played a lesser role than MA. At the same time it must be remembered that, taking the whole range of ontogenetic development into account, MA and CA are closely associated, and this would account for the (chronological) "age-related" sequence in concept development which has been observed in many other studies. In other words, MA and CA are only rarely independent variables, except within a restricted age range, say less than 1 year.

The fact that MA and concept-development scores show only a moderate degree of association does not appear to be simply the result of low test reliabilities or of the use of a comparatively narrow age range. Hood (1962) has noted that some of his Ss, both normal and intellectually retarded, showed a marked discrepancy between MA and expected level of number-concept development, and that in no case did a child at a low stage of conceptual development reveal high arithmetic ability,

TABLE 3

REGRESSION ANALYSIS OF VARIANCE OF ARITHMETIC ATTAINMENT

Source	Arithmetic computation	Arithmetic problems
MA	.295	.310
Concept scores added	.080	.109
F	16.00***	21.80***
Residual	.625	.581

*** $p < .001$.

no matter what his MA. The present results are in agreement with this finding.

It could be argued that the comparatively moderate degree of association between MA and conceptual development results from the fairly low "conceptual" loading of the PMA Test. The same comment could be made regarding most of the group tests of general intellectual ability in common use at the present time, however, and this may well indicate a deficiency which should be remedied.

It appears, then, that while the patterns of development of conceptual thinking and of general intellectual ability may be similar, these patterns do not necessarily coincide. Moreover, children's school performance seems to be associated with aspects of conceptual thinking which are not adequately assessed by conventional intelligence tests. Without further exploration it is not possible to say definitely whether faster concept development contributes to higher school attainment or is a reflection of it, or, in fact, whether both are the observed result of the operation of some third factor. But it can be stated, with assurance, that this relationship will repay further detailed study, and that it may well be one of the keys to future improvements in teaching programs.

REFERENCES

BEARD, R. M. The nature and development of concepts: I. *Educational Review,* 1960, **12,** 157–176. (a)

BEARD, R. M. The nature and development of concepts: II. *Educational Review,* 1960, **13,** 12–26. (b)

BRAUN, J. S. Relation between concept formation ability and reading achievement at three developmental levels. *Child Development,* 1963, **34,** 675–682.

BRUNER, J. S. *The process of education.* Cambridge: Harvard University Press, 1960.

CARPENTER, T. E. A pilot study for a quantitative investigation of Jean Piaget's original work on concept formation. *Educational Review,* 1955, **7,** 142–149.

CHURCHILL, E. M. The number concepts of the young child: I. *Researches and Studies,* 1958, **17,** 34–49. (a)

CHURCHILL, E. M. The number concepts of the young child: II. *Researches and Studies,* 1958, **18,** 24–46. (b)

DODWELL, P. C. Children's understanding of number concepts: Characteristics of an individual and of a group test. *Canadian Journal of Psychology,* 1961, **15,** 29–36.

ELKIND, D. The development of quantitative thinking: A systematic replication of Piaget's studies. *Journal of Genetic Psychology,* 1961, **98,** 37–46.

FEIGENBAUM, K. D. Task complexity and IQ as variables in Piaget's problem of conservation. *Child Development,* 1963, **34,** 423–432.

FLAVELL, J. H. *The developmental psychology of Jean Piaget.* New York: Van Nostrand, 1963.

FREYBERG, P. S. Some aspects of intellectual development in children aged six to nine years: A longitudinal study. Unpublished doctoral dissertation, Victoria University, 1964.

HOOD, H. B. An experimental study of Piaget's theory of the development of number in children. *British Journal of Psychology,* 1962, **53,** 273–286.

HUNT, J. McV. *Intelligence and experience.* New York: Ronald Press, 1961.

PIAGET, J. *The psychology of intelligence.* New York: Harcourt, Brace, 1950.

PIAGET, J. *The child's conception of number.* New York: Humanities Press, 1952. (a)

PIAGET, J. *The origins of intelligence in children.* New York: International Universities Press, 1952. (b)

(Received July 9, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 3, 169-173

GENERALIZED EFFECTS OF PRAISE AND REPROOF

HARRY E. ANDERSON, JR., WILLIAM F. WHITE, AND JAMES A. WASH
University of Georgia

Praise and reproof on a classroom achievement test were administered to 52 university female students, both high and low achievers. The effects of praise and reproof were evaluated in terms of performance on an independent variable. Results indicate that praise has more generalized positive effects on performance than reproof.

Several studies have been conducted to assess the effects of praise and reproof on learning and performance. Schmidt (1941) obtained differential effects between performance and reinforcement, while Child and Whiting (1959) emphasized the importance of level of aspiration as a factor in success and failure. Hurlock's (1925) results indicated that praise was clearly the more effective condition for improving the performance of both high-achieving and low-achieving students, yet an obvious inference from Van De Riet's (1964) data was that academically low students would perform better when reproved and worse when praised, while normal or high-achieving students would perform better when praised and suffer lower achievement when reproved.

Other investigations (e.g., Stevenson, 1961; Stevenson, Keen, & Knights, 1963) had common characteristics that suggested limited application of the results for classroom procedures (cf. Hartup, 1965). Assessment of the effects of praise and reproof has always been measured on a specific behavioral task with no indication of the generalized effects of these kinds of reinforcement. Motivation toward the experimental tasks in these earlier studies appears to have been seriously limited by the reduction of ego involvement in such tasks as tossing marbles or learning nonsense material. Another characteristic of prior investigations in praise and reproof focused on the individual administration of the criterion variable to each subject (*S*); since the experimenter (*E*), alone, was present with each *S*, no dyadic or triadic relationship could contribute to the variance of the praise and reproof reinforcement. Hurlock (1925) differed from other investigators in that the classroom became the focus of her research. Nevertheless, the experimental variable was an arithmetic exercise, and the children were brought individually to the front of the classroom for praise and reproof; the effects were specifically measured by repeated performance with the same variable.

The design of the present study departs from previous research in this area: Primarily, the praise-reproof variable was manipulated in the context of a locally designed achievement test covering the content material in a university-level course. Praise and reproof were given in the form of extreme scores and specially prepared remarks by the professor. Second, the achievement test was administered after the students had spent some time in the course so that some foundation was available to the student for assessing related levels of academic expectancy, or aspiration. Finally, the effects of praise and reproof were measured in terms of a variable entirely independent of the achievement test so that inferences could be made regarding the generalized effects of praise and reproof.

From Hurlock's (1925) and Van De Riet's (1964) results, two hypotheses were formulated for the present experiment. The first hypothesis corresponds with inferences from Hurlock's study: H_1—Praised students will perform better than reproved students. The second hypothesis corresponds to inferences made from Van De Riet's study: H_2—Reproved low-achieving students and praised high-achieving students will perform better than praised low achievers and reproved high achievers. A third hypothesis was also formulated utilizing Van De Riet's results within a framework of the level of aspiration (e.g., Child & Whiting, 1959): H_3—Praised students who estimate that their achievement scores will be relatively high and reproved students who estimate their achievement scores will be relatively low will perform significantly better than reproved students who estimate their scores to be relatively high and praised students who estimate their achievement to be relatively low.

METHOD

Variables

Two tests were used in the experiment. The praise-reproof variable was an objective 4-choice, 75-item achievement (ACH) test based on material presented in a 3-week set of lectures related to evaluation and measurement. On a larger sample of 136 cases, the ACH test had a split-half reliability of .63. An item appended to the ACH test required the students to estimate whether their ACH score would be in the lower, middle, or upper third of the class. The second experimental variable, used to assess the generalized effects of praise and reproof, was a 120-item multiple-choice test (COMP) of simple addition and subtraction of single-digit numbers, with all subtraction differences being positive.

Subjects

The sample consisted of junior and senior university students enrolled in three sections of an introductory course of educational psychology; one male professor taught two sections, while another male professor taught the remaining section. The selection of students for experimental analyses was based on several criteria. In a deliberate effort to obtain distinctly high and low achievers, the data for analyses are based on 52 females, selected from the total number of 128 students, who were present for *all* procedures and who scored either in the upper or lower third of the total ACH-test score distribution; absences, either from one of the COMP or the ACH-test administrations, reduced the sample size. Moreover, since a limited number of males were represented in the upper and lower thirds of the ACH score distribution (9 and 12, respectively), and ample evidence of possible sex differences had been reported (e.g., Schmidt, 1941; Stevenson, 1961), males were eliminated from experimental analyses.

Procedures

The students were administered the COMP in a 4-minute session, under the auspices of a math-test standardization program, 1 day prior to the administration of the ACH test. The COMP test administration was conducted by a "neutral" faculty member relatively unknown to the students. Four days after the ACH test, the students were given sham ACH scores. All females designated for the praise condition in the experiment were assigned random sham scores greater than 65, and those designated for the reproof condition in the experiment were assigned random sham scores of less than 20. The sham scores were written on the inside of slips of paper that were folded once and stapled. The student's name was written on the outside, and the students were told periodically to refrain from any discussion, particularly of test scores. After all scores were handed out, the professor made prepared remarks emphasizing that two or three students had scored much higher (above 60) than anyone had scored in more than 5 years of administering this test, while one or two others had scored well below chance (below 20): "...so you can see, I am at the same time very elated and somewhat embarrassed for a few of you in this class." Immediately after the professor's remarks, the COMP was administered again to the students by the neutral faculty member. Following the completion of all testing, the purpose and details of the experiment were explained to the students.

Analyses

The analyses and inferences are based on incremental percentages of COMP scores. The sample sizes in the eight cells of the experiment are much too small for individual comparisons. The three hypotheses, however, were planned for various groupings of the cells in orthogonal contrasts (e.g., Edwards, 1963, pp. 140–144).

RESULTS

The pre- and post-COMP scores correlated with ACH scores, respectively, at .07 and −.09. The correlation between pre- and post-COMP scores was .96. The COMP, then, may be regarded as a highly reliable variable independent of the ACH test. Moreover, the pre- and post-COMP means were, respectively, 60.31 and 66.94, while the corresponding standard deviations were 10.11 and 9.86.

The two sections taught by one professor had an ACH-test mean of 40.85 and a standard deviation of 7.25; the corresponding statistics for the remaining section were 41.65 and 6.60. The two means did not differ significantly ($t = .606$), so the upper- and lower-third groups were determined on the ACH scores for all 128 students. The upper- and lower-third groups had means, respectively, of 46.71 and 35.11, and corresponding standard deviations of 8.38 and 8.88; the means were significantly different with $t = 14.4894$ ($p < .001$).

The incremental percentage COMP scores for the various groups are presented in Table 1. A conventional Bartlett's test on the within-cell variances yielded a χ^2 of 9.8455 ($p < .30$), indicating homogeneity among variances. While the sample sizes are not exactly proportional, the *N*s did not differ significantly from the average of 6.5 ($\chi^2 = 7.6923$; $p < .50$), so the *N* of 6.5 was used in an approximation method of analysis of variance (Johnson, 1949, pp. 261–266) and in the orthogonal comparisons.

TABLE 1

SAMPLE SIZES AND PERCENTAGE COMP SCORE MEANS AND STANDARD DEVIATIONS FOR GROUPS

Independent variable	Type of data	Groups			
		Low ACH with low ACH estimate	Low ACH with high ACH estimate	High ACH with low ACH estimate	High ACH with high ACH estimate
Praise	*N*	5	5	11	6
	M	114.40	122.80	110.91	118.33
	SD	13.5	13.7	10.8	7.7
Reproof	*N*	4	4	10	7
	M	107.00	106.50	108.60	108.14
	SD	5.5	11.4	7.8	11.4

The analysis of variance for among-cell variation resulted in an *F* value of 1.7067 which was not significant ($p > .05$), but this does not vitiate the planned orthogonal comparisons. The standard error for each of the three contrasts is:

$$\sqrt{(8/6.5)(133.224)} = 12.81,$$

and each *t* value is evaluated with 44 degrees of freedom. Using the data in Table 1, the test of significance for the first hypothesis, that the praised groups are higher than the reproved groups, resulted in $t = 2.8259$ ($p < .001$), so the first hypothesis is accepted; the generalized effect of praise as compared to reproof is a greater increment in performance. For the second hypothesis, that praised high and reproved low ACH groups will perform better than praised low and reproved high ACH groups, $t = -.8743$, which is neither significant nor even in the predicted direction, leading to rejection of the hypothesis. With regard to estimated ACH scores, the third hypothesis states that the praised estimated-high-ACH groups and the reproved estimated-low-ACH groups will perform better than the reproved estimated-high-ACH groups and the praised estimated-low-ACH groups. For the third hypothesis, $t = 1.3099$, which, while not significant ($.10 > p > .05$), is at least in the predicted direction; the third hypothesis, however, is rejected.

Discussion

The present study represents an attempt to measure the generalized effects of praise and reproof on a cognitive variable within the classroom context. A greater increment in performance was realized under praise than under reproof. The results suggest that praise and reproof in one subject-matter area in the classroom may well transfer, in terms of effects on performance, to other classroom behavior that is entirely distinct and independent. In agreement with conclusions of Hurlock (1925), present data supported the traditional credence that praise of student behavior will generally improve performance more than methods of reproof.

Van De Riet (1964) obtained an actual reduction in performance for praised underachievers and reproved normal achievers, yet no such reduction was obtained in the present study, but there may be several reasons for this difference. The samples in her study differ from those of the present study in terms of age, sex, and achievement classification; the samples herein are merely defined as relatively high and low achievers with respect to a single classroom variable. Although the effects of praise and reproof were measured in both studies, Van De Riet focused on task-specific effects, while the present investigation was directed toward generalized effects. Moreover, for the present study, students in all of the classes were "contaminated" by the experimental procedure so there was no way to estimate the average increment in percentage of COMP scores without the intervening praise-reproof variable, and only relative comparisons are possible in the present study.

The level of academic expectancy was a unique variable that had not been considered in many previous studies of praise and reproof. Success and failure, as might be constituted in praise and reproof, are generally recognized as having differential effects (e.g., Child & Whiting, 1959), and Kunst (1959) has even postulated that low achievers have a "need" to fail. The results of the present study indicate that the level of academic expectancy may be an important factor in terms of reinforcement and performance. From an examination of the COMP means in Table 1, it is apparent that praise and reproof have the largest differential effects on those students (both high and low achievers) who estimated that they had maintained relatively high academic performance on the achievement test. Such differential effects, however, were not so pronounced for students who estimated their performances to be relatively low. These differences in performance were not significant in the present study, but a more thorough study of such differences might prove profitable in future research.

REFERENCES

Child, I. L., & Whiting, W. M. Determinants of level of aspiration: Evidence from everyday life. In M. R. Loree (Ed.), *Educational psychology*. New York: Ronald Press, 1959. Pp. 235–247.

Edwards, A. L. *Experimental design in psychological research*. New York: Holt, Rinehart & Winston, 1963.

Hartup, W. W. Social behavior of children. *Review of Educational Research*, 1965, **35**, 122–129.

Hurlock, E. B. An evaluation of certain incentives used in school work. *Journal of Educational Psychology,* 1925, **16,** 145–159.

Johnson, P. O. *Statistical methods in research.* Englewood Cliffs, N. J.: Prentice-Hall, 1949.

Kunst, M. S. Learning disabilities: Their dynamics and treatment. *Social Work,* 1959, **4,** 95–101.

Schmidt, H. O. The effects of praise and blame as incentives to learning. *Psychological Monographs,* 1941, **53**(3, Whole No. 240).

Stevenson, H. W. Social reinforcement with children as a function of CA, sex of *E*, and sex of *S*. *Journal of Abnormal and Social Psychology,* 1961, **63,** 147–154.

Stevenson, H. W., Keen, R., & Knights, R. M. Parents and strangers as reinforcing agents for children's performance. *Journal of Abnormal and Social Psychology,* 1963, **67,** 183–186.

Van De Riet, H. Effects of praise and reproof on paired-associate learning in educationally retarded children. *Journal of Educational Psychology,* 1964, **55,** 139–143.

(Received July 13, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 3, 174–178

EMOTIONALLY LOADED WORDS AND THE ACQUISITION OF A SIGHT VOCABULARY

DAVID R. OLSON[1] AND A. S. PAU
Dalhousie University

A group of 33 primary children were taught to read a list of 6 words, 3 of which were empirically classed as high-emotional words, and 3 classed as low-emotional words, in 2 independent experiments. The subgroups of words were matched for length, frequency, and grammatical function. Training was done by the anticipation method with correction after each trial. Children required significantly fewer trials to learn the high-emotional words. This finding is predictable both from psychological theory and from the observations of insightful teachers such as Ashton-Warner.

An observation of some significance to education has been made by Sylvia Ashton-Warner (1963) in her recent book, *Teacher*. She observed that Maori school children were often completely impervious to the static *Janet and John* reading program. She found that she was able to solve the reading problem by selecting for each child a set of words of intense feeling and meaning; they quickly learned to read such words as ghost, fight, kiss, and mommy, presumably because they reflected the basic needs, fears, and concerns of the children.

Her views have a strong affinity with the writings of John Dewey on the role and nature of interest in education. Dewey (1913) argued that it is impossible to first select subject matter and later make it interesting. Subject matter must be selected in terms of the child's urgent needs and interests.

The purpose of the present study was to recast this view into a psychological model and to verify it empirically. What are the effects of emotional loading on the acquisition of a sight vocabulary? This problem may be considered in terms of the interrelationship between motivation and cognition. Extensive critical reviews of the experimental studies concerning this relationship have been presented by Allport (1955), Chaplin and Krawiec (1960), and Henle (1955). Representative of the studies reviewed was that of Bruner and Postman (1948) who had subjects (*S*s) judge the sizes of disks inscribed with either positive, negative, or neutral symbols. The positive-valued symbol (a dollar sign) was judged larger than the negative-valued symbol (a swastika), but both of these were judged as larger than the neutral one. They concluded that whatever is important to *S*, whether positively or negatively, was likely to be exaggerated in his perception. In a concept-identification situation, Olson and Harper (1963) presented evidence that children do not respond exclusively on the basis of cognitive input; the reactions were modified by the motivational significance of the material to be learned. Hence, in different contexts, emotional significance or affect has either a positive or a negative influence—it is rarely, however, irrelevant.

Another line of evidence comes from the studies of factors affecting the formation of verbal associations. Meaningfulness has repeatedly been shown to facilitate this process.[2] Recently, Koen (1962) has shown that emotionality is a factor independent of meaningfulness as defined by association value. His finding is in line with the position developed in this study. Koen goes on to speculate that emotionality may act as "noise in a communication channel," a speculation totally at variance with the other studies summarized here.

---

[1] Now at the Center for Cognitive Studies, Harvard University.

[2] For example, Palermo, Flamer, and Jen-

For the purpose of this paper, the theory may be stated thus: Except perhaps on occasions of extreme stress, the affective or emotional loading of a stimulus makes it more accessible to the organism by providing a set or expectation. As Henle (1955, p. 426) stated, "It seems that, under the influence of a need or attitude, we are attuned to events to which we would not otherwise be sensitive." This set will have a positive influence when the cue activated is central to the task, and a negative influence if attention to this cue detracts one from the genuine purpose of the task. Since, in our study, any vigilance toward the emotionally loaded material was relevant to the task, the prediction followed that emotional loading facilitates learning. More specifically, it was hypothesized that kindergarten children learn to read words which have an emotional significance more quickly than neutral or nonemotional words.

METHOD

Design

The overall design of this study involved teaching a group of primary school children to associate a phonological reading response with each of a set of visually presented words, one-half of which were Emotionally Loaded and the remaining one-half Nonemotionally Loaded or Neutral. The method employed was the anticipation method with correction after each response.

Subjects

*S*s for this investigation were 9 boys and 9 girls from the Primary Grade at LeMarchant School, Halifax, Nova Scotia. All children were either 5 or 6 years of age, and none of them could read any of the words used in the experiment.

Materials

The materials used in this experiment consisted of a list of words which were classified as either Emotionally Loaded (E) or Nonemotionally (N-E) Loaded. In constructing this list, 46 nouns, verbs, and adjectives were selected from the Thorndike-Lorge (1944) word list with the restrictions that they be words having frequency ratings of 100 or more per million, and suitable for children at the primary level. These 46 words were presented to 30 sixth-grade children who were to classify them as: (*a*) "Words I love"; (*b*) "Words I hate"; or (*c*) "Words that don't matter." Only those words which consistently fell into the same category were used in the experiment. Words in the first two categories were classified as E words, and those in the third as N-E. The categorization of these words was then cross-validated by exposing them to a small group of first-grade children who were to sort these into the same three classes. This procedure duplicated the results obtained by the earlier method.

On the basis of pilot trials, it was concluded that a list of six four-letter words was appropriate for primary children. The words were printed in block letters on 3 × 5 cards. The order of the words was randomized; further, to minimize the effects of order in the children's learning, two sets of cards were used, each set randomized independently and combined to make a set which could then be presented in a cyclic fashion to the children. The words utilized in this study along with the percentage of children ($N = 30$) who rated them as

---

kins (1964) showed that high-meaningful nonsense syllables were learned in a paired-associate task more easily than low-meaningful nonsense syllables for both fifth-grade children and adults. Although the apparent contradiction is somewhat bothersome, meaningfulness is obviously relevant to the present study. The definition of meaningfulness in terms of the number of associations given per minute provides the curious possibility that the meaningfulness of a word can be increased by paying *S*s on a "piece" basis, rather than on an hourly basis.

TABLE 1

TRIALS REQUIRED FOR LEARNING OF EMOTIONAL AND NONEMOTIONAL WORDS

Class & word	Emotional rating	Trials		t
		M	SD	
Emotional				
GOLD	90%	3.6		
BURN	93%	3.9		
KILL	96%	4.4		
Combined mean		3.7		
Total		11.9	3.61	
Nonemotional				
HIGH	13%	6.0		
MANY	33%	5.6		5.09*
HALL	36%	8.8		
Combined mean		6.8		
Total		20.4	6.38	

* $p < .001$, $df = 15$.

Emotional are presented in the first two columns of Table 1.

Procedure

The 18 children employed in the study were individually tested by the experimenter (*E*) in a private room in the school. As soon as a reasonable degree of rapport was established, the child was seated at a table facing *E*. *E* presented the word cards, which had been randomly ordered, one at a time in the palm of his hand. As each word was presented the child was asked to read the word; if he could not, *E* pronounced the word and requested that the child repeat it. *E* simultaneously recorded the reading response (right or wrong) on a form provided. The first time through the list of words, *E* also asked if the child knew what the word meant. Virtually all the children used in this study could use the words in sentences; on the occasions when they could not, *E* provided an explanation until the child signified some understanding.

Words were presented in a cyclic fashion with no stops between the presentations of the list. Learning continued until the criterion of three consecutive correct responses was given to each word in the list. Data for use in the analysis consisted of the number of trials each child required for each word before he responded correctly to it for three or more trials.

RESULTS

Of the results from the 18 *S*s involved in the experiment, only 16 were used in the analysis. Both of the excluded *S*s failed to make any progress on the task before they showed signs of fatigue.

The mean number of trials required for the mastery of each of the six words employed in the experiment with combined means for each of the classes, Emotional and Nonemotional, are presented in Table 1. Since each *S* served as his own control, a *t* test for the significance of the difference between correlated means for the Emotional and Nonemotional words was conducted. Raw scores for this analysis consisted of the total scores of individuals for E words and N-E words. The resulting *t* is also presented in Table 1. An examination of Table 1 shows clearly that the children required significantly fewer trials to learn to read the Emotional words than Nonemotional or Neutral words. There was no apparent difference between the positive and negative words within the Emotional class.

Although the results of this experiment are striking, the factor producing the effect cannot be attributed with absolute certainty to the independent variable, Emotionality, for two reasons. First, it is possible that nouns and verbs are learned more easily than adjectives; the list of words used in the experiment was not controlled for part of speech. Second, two of the words in the Nonemotional class begin with the letter H. It is possible that this redundancy complicated the discrimination of these two words. In order to eliminate these contingencies, the study was replicated.

A REPLICATION

The replication of the experiment was conducted in a manner identical with that already described. The children employed in this study were 15 5- and 6-year-old primary children from St. Thomas Aquinas School in the same area of Halifax as the school in which the first study was conducted. In order to broaden the empirical base of the conclusion, six entirely new words were utilized in the replication. These new words were selected to meet the criteria established in the first section, as well as the limitations mentioned above. The set of words utilized for this replication along with the percentage of children ($N = 30$) who rated them as Emotional are presented in Table 2.

TABLE 2

TRIALS REQUIRED FOR LEARNING OF EMOTIONAL AND NONEMOTIONAL WORDS

Class & word	Emotional rating	Trials *M*	Trials *SD*	*t*
Emotional				
HATE	96%	5.5		
WISH	93%	6.3		
KISS	73%	3.6		
Combined mean		5.1		
Total		15.3	4.33	
Nonemotional				
COME	23%	8.9		4.92*
KEEP	43%	8.5		
BUSY	56%	5.8		
Combined mean		7.7		
Total		23.7	6.64	

* $p < .001$, $df = 14$.

An examination of Table 2 shows that although the words which constitute the experimental classes did not overlap, it was impossible to get words with extreme Emotional and Nonemotional ratings and still meet all the criteria discussed earlier. This difficulty increases the possibility of the Type II error; that is, it staggers the probabilities against the hypothesis in question.

The number of trials required for the mastery of the words for each class is also presented and analyzed in Table 2. This table shows that the mastery of Neutral words required significantly more trials than did the Emotionally Loaded words. These findings confirm the conclusion drawn after the original experiment.

CONCLUSIONS AND IMPLICATIONS

This study has provided empirical evidence that children can learn phonological reading responses to Emotional words more easily than they can to Nonemotional words. In so doing it has provided some evidence for the importance of motivational factors in learning. Furthermore, it substantiates what Ashton-Warner (and other good teachers) have known all along, namely, that the degree of learning is a function of the significance that the material has to the child. The study also indirectly adds to the validity of the concept of intrinsic motivation—motivational factors which arise from within the child. It leaves unanswered, however, the interesting question of the relationship between meaningfulness and emotionality.

The hypothesis tested in this experiment received unexpected confirmation when one *E*, upon entering the school, observed a child emerging from the training administered by the other *E*. *E* asked the child what he had been doing. He replied, "I learned to read some words." When *E* asked what the words were, the child replied, "Kiss. . . , Hate. . . , Wish. . . , and some others that I don't remember." The neutral words were not only more difficult to learn to read, but they were also forgotten more quickly.

Most children in our society, especially in the middle class, are sufficiently motivated by their desire to read that they make satisfactory progress even with relatively neutral material. Nevertheless, highly emotional words may act as a bridge for the nonreader or the poor reader to enter into the reading world. The minimal implication of the study is that teachers and researchers must not consider children as neutral or unstructured intellectual machines passively waiting to be programed. The secret to gaining access to the child's mind is through the use of words and ideas that are important or significant to him, that is, through ideas to which he has already acquired an emotional reaction.

Further studies could be conducted to examine the effects of situational or induced emotionality on learning. Perhaps it is possible to vary the emotional loading of certain concepts by introducing them in a certain way. This would be contrary to the Dewey–Ashton-Warner position that interests must be discovered and exploited and not merely laid on. Such evidence would provide some underpinning for the more orthodox educational practice of beginning a lesson on the prescribed material by writing into a lesson plan a small section called "motivation for the lesson." This study has been an attempt to demonstrate that the relationship between motivation and cognition in an educational context is an important and researchable issue.

REFERENCES

ALLPORT, F. H. *Theories of perception and the concept of structure.* New York: Wiley, 1955.

ASHTON-WARNER, S. *Teacher.* New York: Simon & Schuster, 1963.

BRUNER, J. S., & POSTMAN, L. Symbolic value as an organizing factor in perception. *Journal of Social Psychology,* 1948, **27,** 203–208.

CHAPLIN, J. P., & KRAWIEC, T. S. *Systems and theories of psychology.* New York: Holt, Rinehart & Winston, 1960.

DEWEY, J. *Interest and effort in education.* New York: Houghton Mifflin, 1913.

HENLE, M. Some effects of motivational processes on cognition. *Psychological Review,* 1955, **62,** 423–452.

KOEN, F. Polarization, M, and emotionality in words. *Journal of Verbal Learning and Verbal Behavior,* 1962, **1,** 183–187.

OLSON, D. R., & HARPER, R. J. C. The role of stimulus affect on concept attainment. *Alberta Journal of Educational Research,* 1963, **9,** 125–131.

PALERMO, D. S., FLAMER, G. B., & JENKINS, J. J. Association value of responses in paired-associate learning of children and adults. *Journal of Verbal Learning and Verbal Behavior,* 1964, **3,** 171–175.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words.* New York: Teachers College, Columbia University, Bureau of Publications, 1944.

(Received August 31, 1965)

to help you with student classification

SRA publishes comprehensive survey and placement materials such as the High School Placement Test (HSPT) for grades 8-9, and the Pupil Record of Educational Progress (PREP) for grades 6-9.

THE SRA HIGH SCHOOL PLACEMENT TEST classifies second-semester eighth graders and first-semester ninth graders according to achievement and educational ability. As a guide to the counselor, HSPT helps direct students to the proper high school program (college preparatory, general academic, or commercial), and identifies the gifted as well as those needing extra help.

Subtests include an intelligence test and achievement tests in these areas: Reading, Arithmetic, Modern Mathematics, Language Arts, and Science Methodology.

THE PUPIL RECORD OF EDUCATIONAL PROGRESS contributes an index of a student's achievement level, intelligence, and educational aspirations to your guidance program.

The PREP achievement battery tests these four areas: English, Mathematics, Science, and Social Studies. The IQ portion estimates academic potential, measuring three factors of intelligence: Language, Reasoning, and quantitative abilities. An educational planning questionnaire reveals the student's view of factors influencing his academic future.

Write for a complete listing and description of SRA tests and test services.

SRA

Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A subsidiary of IBM

New and Recent Books in Psychology from

THE DORSEY PRESS

PSYCHOLOGICAL STATISTICS: An Introduction

By *Frederick A. Courts*, Reed College

Designed for first courses in statistics offered in departments of psychology, education, or sociology, this new volume provides a treatment of the underlying rationale and typical applications of statistical methods commonly used in psychology. A minimal mathematical background is required. Extensive use is made of figures and graphic illustrations. *About 400 pages. Just published.*

FOUNDATIONS OF THE THEORY OF PREDICTION

By *William Warren Rozeboom*, University of Alberta

This new volume provides basic text material for courses of one- or two-semester length in test theory. It will also have value as a supplement or reference for courses in tests and measurements or multivariate analysis. The topics presented are basic ones, but they are analyzed in greater depth than anywhere else in the literature of the field. *About 500 pages. Summer 1966.*

READINGS IN PSYCHOLOGICAL TESTS AND MEASUREMENTS

Edited by *W. Leslie Barnette*, State University of New York at Buffalo

Published in both cloth and paperbound editions, this unique volume provides a selection of special readings drawn chiefly from psychological journals and designed to accompany a basic text in tests and measurements. The selection of articles is based on the editor's experience in teaching undergraduate classes. *365 pages. Copyright 1964. Cloth: $5.50. Paper $3.00.*

Write for examination copies for college adoption.

Published in The Dorsey Series in Psychology

Consulting Editors: *Howard F. Hunt*, Columbia University
Donald W. Taylor, Yale University

DORSEY PRESS

THE DORSEY PRESS

HOMEWOOD, ILLINOIS

Three texts of major importance

"The volume is firmly based on sound learning theory and the applications are pertinent and cogent."

THE PSYCHOLOGY OF CLASSROOM LEARNING

J. M. STEPHENS, *University of British Columbia*

"The text has been admirably designed for . . . teachers, regardless of the area and the level of their interest. The volume is firmly based on sound learning theory and the applications are pertinent and cogent. . . . The author is to be commended upon the depth and breadth of the references quoted, most of which date since 1960." William P. Sears, in *Education*, December 1965

1965 **525 pp.** **$8.50**

Provides a comprehensive examination of the role of education in cultivating creativity. . .

THE ART AND SCIENCE OF CREATIVITY

GEORGE F. KNELLER, *University of California, Los Angeles*

Summarizes the current state of our knowledge regarding the creative process . . . places current concepts of creativity in historical, philosophical, and behavioral perspective . . . and cuts across disciplines to consider the common source of creativity in arts and sciences. In addition, the book suggests ways in which creativity can be fostered in our present school systems.

1965 **128 pp.** **$2.50 paper**

"Wallach and Kogan have made significant contributions to conceptual differentiation of creativity and to its measurement. . . ."

Teachers College Record, February 1966

MODES OF THINKING IN YOUNG CHILDREN
A Study of the Creativity-Intelligence Distinction

MICHAEL A. WALLACH, *Duke University*
NATHAN KOGAN, *Educational Testing Service*

"I found the book excellent. It is a well written and elegantly executed piece of research—a fine addition to the work on creativity in children." Jerome Kagan, *Harvard University*

1965 **368 pp.** **$8.75**

Holt, Rinehart and Winston, Inc.
383 Madison Avenue,
New York, New York 10017

Reading in High Gear

Reading in High Gear is a new approach to the problem of teaching reading; this program assumes only that the student can speak English.

Reading in High Gear has been developed by Dr. Myron Woolman, primarily for the culturally disadvantaged adolescent and adult. Dr. Woolman's program is based on his Progressive Choice Learning Theory, first validated in his study "The Effect of Varying the Number of Choices in the Identification of Very Similar Stimuli."*

The two basic principles of Progressive Choice are: Discrimination . . . two things taught in sequence should be dissimilar in appearance to minimize confusion and facilitate learning; Invariance . . . in the initial learning steps, only one response will be required for any one element taught.

Reading in High Gear is particularly appropriate for teachers looking for suitable material to meet individual abilities of remedial and pre-dropout students. For further information about Progressive Choice and **Reading in High Gear,** see the noted articles and return the coupon to SRA.

*1. Woolman, Myron, "The Effect of Varying the Number of Choices in the Identification of Very Similar Stimuli," unpublished doctoral dissertation. Columbia University, 1955. Complete microfilm copy available from University Microfilms, Ann Arbor, Michigan.

2. "New Teaching Technique Raises Reading Level," **Jet,** August 26, 1965.

3. Bloomer, R. H., "Reading Methodology: An Experimental Test of Some Alternative Organizational Principles," **The Reading Teacher,** January 1961.

4. Rothberg, I. C. and Woolman, M., "Verbal Paired-Associate Learning as a Function of Grouping Similar Stimuli or Responses," **Journal of Experimental Psychology,** January 1963. Vol. 65, No. 1.

Please send a free copy of Dr. Woolman's twenty-page booklet about **Reading in High Gear.**

Name____________________

Position____________________

School____________________

Address____________________

City__________State__________Zip__________

S R A
Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A subsidiary of IBM

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 57 August 1966 Number 4

CONTENTS

Effect upon Transfer of Variations in Training Conditions 179
ROBERT L. R. OVERING AND ROBERT M. W. TRAVERS

Audience Recall as a Function of Speaker Dynamism 189
WILLIAM D. COATS AND ULDIS SMIDCHENS

Perceptual Span Training and Reading Achievement of School Children 192
BRUCE R. AMBLE AND SIEGMAR MUEHL

Description of Junior Colleges 207
JAMES M. RICHARDS, JR., LORRAINE M. RAND, AND LEONARD P. RAND

Combining Response Modes in Programmed Instruction 215
JOANNA P. WILLIAMS

Learning and Measured Abilities 220
JAMES P. DUNCANSON

Ability to Match Information within and between Auditory and Visual Sense Modalities and Subsequent Reading Achievement 230
SIEGMAR MUEHL AND SHIRLEY KREMENAK

Journal of Educational Psychology
1966, Vol. 57, No. 4, 179–188

EFFECT UPON TRANSFER OF VARIATIONS IN TRAINING CONDITONS[1]

ROBERT L. R. OVERING
McGill University

AND

ROBERT M. W. TRAVERS
Western Michigan University

An experiment was conducted to investigate the effect upon transfer of: (a) varying irrelevant information while holding relevant information and time constant, (b) establishing a set, and (c) verbalizing knowledge acquired during training. 96 Ss, with an IQ range of 110–130 in Grades 6 and 7, were assigned to 4 different conditions of learning. ½ of the Ss in each learning condition were asked to verbalize the knowledge acquired before attempting to solve 2 problems involving aiming at an underwater object. Results suggest that transfer is facilitated by: (a) training in the presence of irrelevant information, (b) establishing a set relating the material to be learned to a practical situation, and (c) verbalization of knowledge prior to its application.

Previous experiments (Hendrickson & Schroeder, 1941; Judd, 1908) indicate that verbal knowledge of a principle has a facilitating effect on the amount of transfer evident in a situation requiring application of that principle. Hendrickson and Schroeder, further, showed that the amount of transfer was related to the amount of information taught, that is, the less information, the less transfer, and vice versa.

Neither of these previous experiments, however, investigated the effect of teaching the principle by different means. In fact, in the experiment reported by Judd, we are not informed at all of the manner in which the information making up the independent variable was taught, or, for that matter, what information was taught. We are simply told that the experimental group was given a full theoretical explanation of refraction (the subjects were Grade 5 and 6 boys), while the control group received no theoretical training. As a test for the effect of this knowledge on transfer, all subjects (*S*s) were then tested in a series of tasks requiring them to throw darts to hit an underwater target.

Hendrickson and Schroeder in their 1941 experiment, which was a modified reproduction of Judd's, do describe both the method by which they taught the information and the information itself. Essentially, each *S* was presented with a brief printed statement and a diagram representing (somewhat inaccurately) the refraction of light in water and the resulting apparent displacement when an underwater object is viewed at an angle from above. Each *S* was permitted to study this information until he stated that he understood it. As in Judd's experiment the control group received no instruction or information relating to refraction or displacement. Like Judd's, also, the transfer task again was to attempt to hit an underwater target, only in this case an air rifle and pellets were substituted for the darts.

The experiment which is described in this report was an attempt to extend the findings of Judd, and of Hendrickson and Schroeder, to see what the effect upon transfer might be of introducing variations in training conditions, where relevant information and time were held constant. The dimension varied was essentially that of irrelevancy; for example, in the "realistic" treatment, all sorts of irrelevant cues were present, whereas in the "visually compressed" treatment an effort was made to eliminate many irrelevant bits of information by reducing the visual aspect of the presentation to a series of line drawings. The choice of treatments arose from a survey of research and theory related to audiovisual information transmission recently completed at the University of Utah by Travers et al. (1964).

---

[1] This article is based on research sponsored by the New Educational Media Branch, United States Office of Education, under Project C-997. Complete details of the procedures described can be found in a doctoral dissertation by R. L. R. Overing which includes this and other studies.

In Travers' monograph it is pointed out that evidence exists to indicate that irrelevant information may interfere with learning (e.g., Hunt, 1962). This viewpoint is backed by neurophysiological and other evidence which suggests that the human senses and nervous system compress information into simpler forms (Attneave, 1954; Cherry, 1962; Hubel, 1963). Thus the transmission of information should take place most effectively under conditions where the information to be transmitted has been precompressed, either by the elimination of redundancy and/or the elimination of irrelevant information.

However, the possibility also exists that conditions of compression, while facilitating transmission of information, may not be optimum where transfer is called for. In other words, the optimum conditions for learning a concept may *not* be the optimum conditions for transfer related to that concept. Our research, the main details of which are reported below, is an attempt to cast some light on these problems.

On the presumption that S's knowledge of a principle tends to raise S's level of performance in a transfer task requiring application of that principle, the experimenters set out to determine:

1. Whether knowledge of the principle of refraction acquired under visually compressed or under realistic training conditions would best facilitate transfer.

2. Whether verbalization of the knowledge taught would facilitate performance on the transfer task.

3. Whether contact with a relevant problem immediately preceding instruction with visually compressed materials would establish a set, thus facilitating performance on the transfer task.

METHOD

Experimental Design

The experimental design was a 2 × 2 × 4 × 2 × 12 factorial design comprising a Verbalization-Nonverbalization condition, 2 sexes, 4 training conditions, 2 tests, and 12 repeated measures within each test.

Training Conditions

Four training conditions were introduced into the experiment. The attempt was made to teach all groups the same information but to vary the way in which the information was taught. The time devoted to teaching was the same for all groups. The verbal information provided was reproduced in each case by means of a tape recorder. The four teaching conditions were as follows:

1. *Verbal Information Only.* In this condition Ss listened to a tape which provided all the information they received. The script described: (*a*) the principle of the refraction of a light ray entering and leaving water, and (*b*) the phenomenon of the visual displacement of an underwater object when viewed at an angle from above. The time taken to run this tape became the base time used in all teaching procedures. The script of the tape also became the basis for verbal materials used in all other teaching conditions.

2. *Visually Compressed Information plus Verbal Information.* In this teaching condition line drawings were introduced to transmit some of the information and some of the redundant items of the original verbal script were deleted. The drawings pro-

vided very simple representations of the phenomena discussed, but colors were used to facilitate the identification of parts of the drawings from the verbal descriptions. Every effort was made to ensure that the information content of this and all other teaching procedures was identical.

3. *Realistic Demonstration plus Verbal Information.* This teaching condition involved a visual presentation of a light beam passing into the water in a 2-gallon fish tank. The beam was reflected by a mirror on the bottom of the tank and also could be observed as it left the water. The beam of light was produced by a 500-watt filmstrip projector arranged to produce an approximately parallel beam of light which was .25 inch in diameter at the place where it could first be observed. In order to improve the visibility of the beam of light, chalk dust was knocked out of a blackboard eraser over the tank. In this condition the pupils were free to observe the beam of light from different angles.

In addition, pupils were given a demonstration of the phenomenon of the visual displacement of an underwater object. This involved two identical tanks side by side, and *S*s were shown that when water was added to one of the tanks, the apparent position of a coin on the bottom appears to move upwards and also away from them. Pupils were shown how the position of a coin in the parallel empty tank had to be changed to make it appear to be in the same relative position as a coin in the similar tank partly filled with water.

4. *Visually Compressed Information plus Verbal Information Preceded by an Orienting Problem.* This teaching condition was the same as Condition 2, but pupils were first given a problem to point out some of the practical implications of the principle of refraction. In this problem the pupils were told the story of an explorer who had to cross a river, but who first had to shoot and kill a crocodile 3 feet below the surface and 20 feet from the explorer. The pupils were asked to indicate on a diagram the point on the surface of the water where the explorer should aim. Pupils were not told whether their solutions were right or wrong, but the purpose was only to focus attention on a practical problem involving the phenomenon of refraction.

Half of the *S*s of each training condition were required to undertake a written test in which their task was to verbalize aspects of the information given. The other half of the *S*s proceeded directly to the testing situations.

In the case of some of the groups of *S*s run early in the experiment, every fifth *S* was assigned to a control group that was given no training, but proceeded immediately to the testing situation. The function of this group was to determine whether training did produce learning.

Test Conditions

The tasks used to measure differences in the amount of transfer resulting from each of the four training conditions involved the activity of shooting at a target under water. The task was similar to that used in the previous studies of Judd (1908) and Hendrickson and Schroeder (1941), but certain important modifications were made. The testing situation involved a mock pistol mounted on a stand in such a way that only the angle of elevation could be varied. The target object was immersed in a large tub of water, and the surface of the water in the tub was covered with a rigid transparent plastic sheet on which was marked a numbered grid. The pupil in the test situation did not actually have to shoot, but was required only to indicate the numbered line on the grid at which he was aiming. This procedure was designed to eliminate differences in the ability to aim and shoot a gun, which would have contributed substantially to the error term in the subsequent analysis.

Each *S* performed two similar tests which differed in that the water was deeper in the second test (9 inches in comparison with 5 inches), and the second test involved a more oblique viewing of the object under water. In the first test the lower edge of the tank was 31 inches horizontally from the gun stand, and the gun itself was 39.5 inches above the floor, but in the second test the horizontal distance was increased to 47.5 inches and the vertical height to 48.75 inches. The second test also involved a different shaped target and tub and was administered in a different room than the first test. It appeared to involve a more difficult task than the first test.

The *S*s were instructed before the tests that each time they reported their point of aim they would be told either "hit," "miss," or "bull's-eye." They were also told that a bull's-eye meant that they had hit the very center of the target, while a hit meant that they had struck some part of the target but not the center area. They were also told the depth of the water for each test. Each *S* was allowed up to 12 shots to find the bull's-eye. Those who hit the bull's-eye before the twelfth shot had the task terminated.

The *S*s were tested individually, and, while each was asked not to give away the answer to other pupils, some effort was made to arrange the procedure so that the pupils returned to their classes after the testing where little opportunity to discuss the experience was provided. In one of the early tryouts of the materials, leakage of information concerning the correct place to aim the gun was identified as a serious problem.

The small group of *S*s run as a control group without any training was given special instructions as follows:

> Underwater objects, when viewed at an angle from above, do not really lie in the exact position that they appear to lie. Knowing this fact, your task, now, is to aim the pistol so that a bullet from it would strike the underwater target.

Subjects

The *S*s ($N = 215$) were fifth, sixth, seventh, and eighth graders drawn from three public schools in Salt Lake City, Utah, as follows:

1. Stewart School (campus demonstration school, K-IX) Grade 8, 40 pupils; Grade 7, 40 pupils; Grade 5–6 (combined grade), 32 pupils.
2. Wasatch School (middle class, K-VI) Grade 6, 39 pupils.
3. Whittier School (lower middle class, K-VII) Grade 7, 32 pupils; Grade 6, 32 pupils.

The Grade 7 and 8 pupils from the Stewart Demonstration School ($N = 80$) served as pilot subjects, and 8 of them provided the control group given no training. Their test results suggested several modifications in the experimental conditions, namely, the importance of controlling for information leakage among *S*s, and the desirability of increasing the difficulty of the second test of transfer. In addition, the pilot group served to show that differential amounts of learning were occurring among training conditions, and that, therefore, the project was worth pursuing.

Of the remaining 135 *S*s, one complete block of 32 was dropped (Whittier School, Grade 6) since their scores indicated that no learning had taken place. Although no definite reason for their chance-level scores can be given, the authors hypothesize that the lack of learning was due to the unsophisticated level of the *S*s, and the high anxiety level which appeared to be generated among them by having to take part in a mysterious activity which involved, among other things, being called down to the principal's office (the rendezvous point), being lined up outside the nurse's office (one of the testing stations), and the anxiety generated by having measures taken of their learning (although in all cases *S*s were assured that their scores would be known to no one in the school and would never form any basis for a report card mark).

Of the remaining 103 *S*s, 7 from Wasatch served to provide some of the control data. The data in Table 1, therefore, are based on 96 *S*s—16 boys and 16 girls at each of three schools.

The 96 *S*s undertook the experiment at their respective schools. At each school *S*s were preselected in the sense that pupils with IQs either below 100 or much above 130 were not considered, and equal numbers of boys and girls were assigned to each training condition. *S*s within each school were then randomly assigned to one of the four training conditions.

RESULTS

The two main measures obtained from each *S* were the scores for Transfer Tests 1 and 2. These scores were

TABLE 1

NUMBER, SEX, AGE AND IQ OF SUBJECTS, BY SCHOOLS

School	Grade	*N*	Sex		Age range	$\bar{X}$ age	IQ range	$\bar{X}$ IQ
			M	F				
Wasatch	6	32	16	16	11-3 to 12-2	11-7	103–133[a]	117.62
Stewart	5–6	32	16	16	11-0 to 12-3	11-7	101–131[b]	113.11
Whittier	7	32	16	16	11-11 to 13-4	12-8	103–127[a]	113.72
Total		96	48	48				

[a] Pintner Intermediate Test.

[b] Wechsler Intelligence Scale for Children.

converted to error scores by subtracting the value of the criterion aiming point from the value of the point of aim for each shot.

Figures 1 and 2 show the learning curves which resulted for Tests 1 and 2. These curves give the impression that no differences between treatments occurred on Test 1, but that differences did occur on Test 2, a finding which corresponds closely with that reported by Judd (1908). On Test 2 the realistic condition and the condition involving compressed visual material preceded by a realistic problem appear superior to the other two. An arbitrary decision was made to run analyses on the scores of the first seven shots only, since scores obtained on shots beyond this point were no longer representative of the group. By shot number seven, most members of most treatments had achieved criterion, and, thus, shots beyond this point tend to be weighted heavily by the scores of one or two *S*s who showed little or no learning. Obviously, too, with a sufficient number of shots, curves of all four treatments would eventually coincide and differences between treatments would disappear.

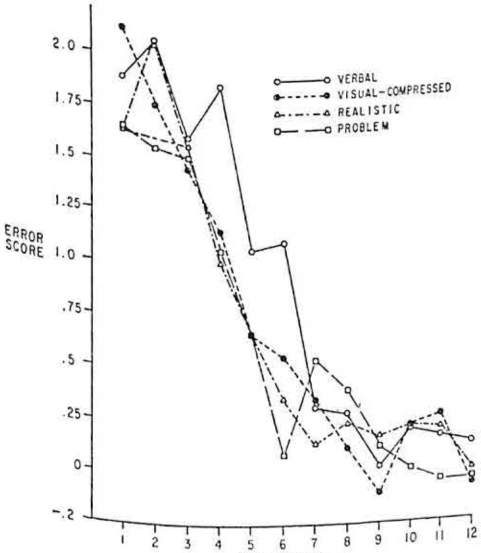


Fig. 1. Test 1 learning curves showing each condition separately.

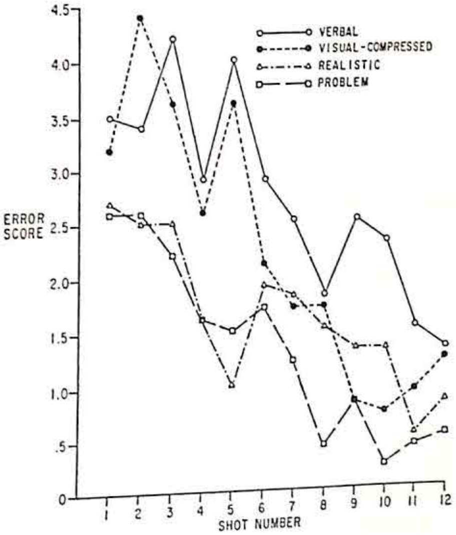


Fig. 2. Test 2 learning curves showing each condition separately.

Control data, used for establishing a base line from which the amount of learning resulting from instruction could be estimated, were obtained from a block of eight control *S*s in the pilot group and a second block of seven *S*s drawn from one of the experimental schools. These *S*s were merely given Tests 1 and 2, and apart from a small amount of information contained in the initial sentence of their instructions for Test 1, they received no instruction or information relating to refraction or displacement. Comparisons were made between the control and the experimental groups using *S*s' mean *error* scores. A significant difference ($p < .01$) was obtained on a t test between the error scores of the control groups and the error scores of the experimentally trained groups, as shown in Table 2. In addition, a second t test was made comparing control and experimental groups, but this time making use of trials to *criterion* scores. The results

TABLE 2
COMPARISON OF CONTROL MEAN WITH THE MEAN OF CONDITIONS 1–4 OBTAINED IN TEST 1

	School	Control mean ($N = 15$)	Combined experimental mean ($N = 64$)	t
Error scores, shots 1–7	Wasatch, Grade 6 Stewart, Grade 7	2.41	1.18	2.90*
Trials to criterion	Wasatch, Grade 6 Stewart, Grade 7	8.45	5.45	3.28*

* $p < .01$, $df = 77$ (2-tailed test).

of this test were also significant ($p <$ .01).

Table 3 gives means and standard deviations for performance on Test 1 and Test 2, respectively, for each of the training groups. An analysis of variance was made of Test 1 error scores, trials 1–7, and also of Test 2 error scores, trials 1–7. The analysis of the results of Test 1 showed no significant differences between treatments, or between those who verbalized and those who did not, or between male and female *S*s.

These results, coupled with those of Table 2 comparing control and experimental groups, might appear to indicate that, although teaching pupils information facilitates performance in a task requiring application of that information for its solution, the manner in which the information is taught is of little consequence. However, the analysis of variance of the scores of the more difficult Test 2 shows a different picture. Here the difference due to treatments results in an $F = 2.64$ ($p < .06$). Although we are aware of the limitations of a t test applied under these conditions, on the basis of our particular interest in Conditions 2 and 3, which vary irrelevant information, a t test was run using Test 2 mean error scores. The results indicate

TABLE 3
MEANS AND STANDARD DEVIATIONS OF MEAN ERROR SCORES FOR PERFORMANCE ON TEST 1 AND TEST 2

Learning condition	Test 1				Test 2			
	Group required to verbalize		Group not required to verbalize		Group required to verbalize		Group not required to verbalize	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Verbal information only	1.60	1.49	1.24	1.70	2.32	2.43	4.34	2.62
2. Visually compressed information plus verbal information	1.36	.80	.99	.64	2.99	1.26	3.19	2.34
3. Realistic demonstration plus verbal information	.62	.76	1.54	1.16	1.45	1.19	2.58	2.79
4. Problem plus Learning Condition 2	.52	.95	1.54	1.55	1.94	2.39	1.87	2.36

Note.—$N = 12$ for each cell.

TABLE 4

SUMMARY OF DUNCAN'S NEW MULTIPLE-RANGE TEST COMPARING MEANS OF TEST 2 ERROR SCORES FOR CONDITIONS 1–4 AND SHOTS 1–7

Treatment	4	3	2	1
M	1.90	2.01	3.09	3.33
Conditions				
Problem (4)	—	*ns*	<.10	<.05
Realistic (3)		—	<.10	.05
Visual-compressed (2)			—	*ns*
Verbal (1)				—

that Training Condition 3, making use of three-dimensional physical apparatus, produced significantly more transfer than did Training Condition 2 ($p < .06$, $df = 46$). Furthermore, application of a Duncan's new multiple-range test indicated that Test 2 scores of Treatments 3 (realistic) and 4 (problem) reflected significantly more learning than did scores on Treatments 1 (verbal) and 2 (visually compressed), at levels varying from $p < .05$ to $p < .10$ (Table 4). In addition, in the Test 2 analysis of variance a Verbalization-Nonverbalization main effect reached the .08 level of significance in favor of the verbalization condition.

DISCUSSION

The results obtained would appear to confirm the findings of Judd (1908) and Hendrickson and Schroeder (1941) that knowledge of a pertinent principle facilitates performance in a task requiring application of that principle. Furthermore, the results obtained from the analysis of variance Test 2 scores would appear to indicate that when a transfer task is sufficiently difficult, the amount of transfer evident is related to the *manner* in which the relevant principle is taught.

The results indicate that one of the significant variations in manner of teaching is the presence or absence of irrelevant information; that is, in terms of performance on Test 2, the realistic condition (3), with all its irrelevant cues, is significantly better than the visually compressed condition (2), which has had most of the irrelevant visual cues removed.

However, we must note that teaching Conditions 2 (visually compressed) and 4 (problem) are identical, except for the introduction of a problem to establish set as the initial phase of Condition 4, and yet Condition 4 results are significantly better than Condition 2 results. It would appear obvious then that it is not the presence of irrelevant information alone which can account for the facilitation of the application of the principle to a new situation. The establishment of a set, prior to dealing with the visually compressed material, appears also to have had the same facilitating effect upon performance as the presence of irrelevant information in the realistic condition (3).

How can the failure of the visually compressed treatment (2) be explained? Probably, on occasion, line drawings compress information in a psychologically inappropriate fashion, or actually eliminate relevant information.

However, even if this were true of the line drawings used here, we are still left with the task of explaining why the mere act of preceding Condition 2 with a rather simple problem for which, and from which, *S*s received no feedback should have had such a facilitating effect. If we assume that some set was established and that somehow some process of selective attention was aroused, we must still ask such questions as:

1. Does the establishment of a set somehow permit *S* to reshape *inappropriately compressed* information into an appropriately compressed form?

2. Does the establishment of a set inevitably provide additional *relevant* information?

Perhaps a clue to the solution of the problem lies in the second of these questions. Berlyne (1960, p. 63) notes that one of the factors affecting processes of selective attention is a class of stimulus events which he labels *indicating stimuli*, a class which includes verbal directions. It is logically conceivable that the establishment of set serves to arouse processes of selective attention similar to those aroused by Berlyne's indicating stimuli. If this were so, possession of an appropriate set might serve to permit *S* to select out from the stimulus complex those stimulus elements which contained the most information. These highly loaded pieces of information to which he has now directed his attention may, in turn, serve to permit *S* to process the remaining information in a manner different from another *S* who, lacking set, perceives the stimulus complex in a relatively unselective fashion. In other words, we are really suggesting that provision of a set may be comparable to the provision of additional information. Reese (1964) has discussed the phenomenon of *performance set* which would also appear to be closely related to the point of view expressed here.

An explanation of the facilitating effect of set, such as the above, however, seems to do little to explain the apparent facilitating effect of irrelevant information, which was evident in the realistic treatment (3). One clue did emerge from the written answers provided by the verbalization task. Several of the written answers of pupils exposed to the conditions which used line drawings stated that the coin appeared to move "to the *right* side, and slightly closer to the surface of the water." *S*s who provided this type of answer appear to have failed to differentiate between those drawings which showed an end elevation and those which showed a side elevation. This confusion was not evident with *S*s in the realistic treatment. In addition, inasmuch as the criterion tests involved a form of depth perception and inasmuch as depth perception is dependent on three dimensions, or the illusion of three dimensions, it is possible to hypothesize that the visually compressed treatment, lacking a third dimension, or even the illusion of a third dimension, actually was deficient in relevant information as compared with the realistic treatment.

The third variation in training conditions investigated in this experiment was provided by the Verbalization-Nonverbalization conditions, wherein one-half of the *S*s of each treatment were required to give written answers to two questions relating to the information to which they had just been exposed. No significant differences between the Verbalization-Nonverbalization groups appeared in Test 1. In Test 2, a difference favoring the group who verbalized did appear, but was significant only at the .08 level. This, however, would suggest the possibility that verbalization prior to a transfer task of sufficient difficulty may facilitate performance.

Finally, we must ask why transfer was very much the same for all the experimental training conditions in the initial transfer test. Judd, himself, noted a related phenomenon in his original experiment and attributed it to the fact that his *S*s were still learning to throw the darts. As very little learning of an "aiming" nature was required in this present experiment, his explanation would seem not to apply. Hendrikson and Schroeder concluded from their experiment that possession of knowledge did facilitate transfer in the *initial* task. However, an examination of the results on which they base this conclusion shows that, by present day standards, their differences between the control group and the two experimental groups, or between the two experimental groups themselves, were not statistically significant.

Throughout the discussion, to this point, we have observed that certain variations in training conditions appear to have a notable facilitating effect on performance of a more *difficult* second task. If we are correct, and if we find no facilitating effect of these same variations in training conditions upon a similar but less difficult task, we would appear to be justified in concluding the following: *S*s in *all* experimental treatments learned enough to solve the relatively easy Transfer Task 1; that is, for its solution it required relatively little learning, and this minimal amount of learning was provided for by all four treatments. In contrast, Test 2, being considerably more difficult, required a level of learning for its solution which had been acquired only by those *S*s exposed to Conditions 3 (realistic) and 4 (problem).

However, at least two other possibilities may be considered. The first of these is that the scores on Test 1 were at the chance level for *all* treatments. That is, having fired at the visual target and being told that it was a miss, as many *S*s would then proceed to lower their point of aim as would proceed to raise it. As aiming below the apparent target as little as one point of the scale on the grid would result in a hit, those who lowered their aim would soon be on their way to a bull's-eye. Although the comparison of control scores with treatment scores shown in Table 2 would seem to discount this possibility of chance-level performance, we must recognize that control data for this experiment is somewhat deficient both in *N* and in comparability.

The second possibility is related both to the *difficulty* of Test 2 and to its *position*. While performances in Test 1 were essentially the same for all experimental groups, only *S*s of Conditions 3 (realistic) and 4 (problem) were able to benefit substantially from the *learning experience* which resulted from participation in Test 1. That is, we might hypothesize that, because of their exposure to a particular type of instruction, *S*s were able to process the feedback which they received while participating in Test 1 differently from those *S*s who had received a different type of instruction. Thus, while all *S*s, in effect, went into Test 1 with equal ability, those who had earlier received instruction in Conditions 3 and 4 emerged from Test 1 with superior ability.

In the light of the results reported and discussed above, it would appear that we may conclude tentatively that the most efficient conditions for transfer exist where irrelevant cues present in the testing conditions are also present in the training conditions. To generalize further, we might hypothesize that the most effective learning, as regards transfer, occurs in the presence of irrelevant cues. To test this hypothesis we must verify whether the negative effect of the visually compressed treatment was due to the particular line drawings used, or whether it was due to the absence of irrelevant cues per se. Second, we must ask ourselves what the relationship is between the presence or absence of irrelevant cues in the training situation and the presence or absence of irrelevant cues in the test situation. To answer these questions an experiment was designed which will be reported on separately.

REFERENCES

ATTNEAVE, F. Some informational aspects of visual perception. *Psychological Review*, 1954, **61,** 183–193.

BERLYNE, D. E. *Conflict, arousal and curiosity*. New York: McGraw-Hill, 1960.

CHERRY, C. E. The band width compression of television and facsimile. *The Television Society Journal*, 1962, **10,** 40–49.

HENDRICKSON, G., & SCHROEDER, W. H. Transfer of training in learning to hit a submerged target. *Journal of Educational Psychology*, 1941, **32,** 205–213.

HUBEL, D. H. The visual cortex of the brain. *Scientific American*, 1963, **209,** 54–62.

HUNT, E. B. *Concept learning*. New York: Wiley, 1962.

JUDD, C. H. The relation of special training to general intelligence. *Educational Review*, 1908, **36,** 28–42.

REESE, H. W. Discrimination learning set in children. In L. P. Lipsitt & C. C. Spiker, (Eds.), *Advances in child development and behavior*. Vol. 1. New York: Academic Press, 1964. Pp. 115–145.

TRAVERS, R. M. W., ET AL. Research and theory related to audiovisual information transmission. Interim Report, July 1964, Contract No. 3-20-003, United States Department of Health, Education and Welfare, Office of Education.

(Received September 8, 1965)

As of September 1, 1966, all manuscripts should be sent to the incoming Editor, Wayne H. Holtzman, Office of the Dean of the College of Education, University of Texas, Austin, Texas 78712.

Journal of Educational Psychology
1966, Vol. 57, No. 4, 189-191

AUDIENCE RECALL AS A FUNCTION OF SPEAKER DYNAMISM

WILLIAM D. COATS AND ULDIS SMIDCHENS
School of Education, University of Michigan

Various relationships among speaker dynamism, grade threat, individual speaker, and immediate recall with respect to a sample of 184 college speech students were explored. The data were analyzed by means of a 3-factor, fixed-effects analysis of variance and the ω^2 test for strength of association. The only relation which was both significant and strong was between speaker dynamism and immediate recall of the speech by the audience. The relation produced a significant F ratio at the .001 level and had a strength of association (ω^2) of .36. It was concluded that audiences remember more from a dynamic lecture than from a static lecture.

Irrespective of whether or not the lecture method is the best technique for teaching, it is in fact used extensively in many classrooms and for students of all ages. Hence, new investigations of delivery variations, which might influence lecturer effectiveness, are always needed. The particular aspect of lecturing in which the authors of this paper were interested was that of auditor attention. Attention was defined as the process of bringing the sense organs to bear upon some subset of stimuli out of the many available in one's perceptual field.

The theoretical development of the study was built upon the following three assumptions: (*a*) Attention is selective (an individual pays more attention to some stimuli than to others); (*b*) attention is fluctuating (the duration of absolute attention to a single stimulus is only a few seconds); (*c*) in order to remember directly any statements made by a lecturer one must attend to some part of the lecture. The experimenters believed these assumptions were justified by intuitive and logical considerations as well as by physiological and psychological research (Billings, 1914; Chapman & Brown, 1935; Pillsbury, 1908, Chs. 1–4).

Of the many concepts which relate to the problem of attention, the authors felt that dynamism (change or variety, animation, and power in the speaker) is the most important. An examination of the three above assumptions indicated that a dynamic lecture delivery should have more success than a static one in eliciting and maintaining the attention of a class. The specific hypothesis follows:

1. Students will remember more from a given lecture when the delivery is dynamic than when it is static. This was the hypothesis of major interest around which most of the design of the experiment evolved.

Furthermore, it was believed that artificial motivation due to grade threat might positively affect class attention. Although such a consideration was somewhat divergent from the main thrust of this paper, relevant data were obtainable with little change in the experimental design. The second hypothesis follows:

2. Students who are told that they will be tested on the content of a lecture will remember more from that lecture than students who do not receive such a grade threat.

METHOD

Design

The statistical design consisted of a three-factor, fixed-effects analysis of variance. One treatment variable was type of lecture, another was grade threat, and the third was the individual doing the lecturing. The dependent variable was the immediate recall by the class of the lecture. Also, the design called for the calculation of the various ω^2 (Hays, 1964) values in order to ascertain the proportion of variance accounted for by the respective treatment effects. A planned-comparison technique was used to evaluate audience perceptions of speaker dynamism under the various treatment combinations.

Procedure

The subjects (*S*s) consisted of 184 students in eight introductory speech classes at the University of Michigan. Two speakers, graduate students in education at the University, each lectured to four different classes. All speeches were about classroom-interaction analysis, were expository in nature, had identical content, and lasted approximately 10 minutes.

Of the four speeches given by each speaker, two were delivered with a dynamic presentation and two with a static presentation. The static speaker read the entire speech from a manuscript. He made no gestures or direct eye contact and held vocal inflection to a minimum. However, he did speak with good diction and sufficient volume. The dynamic speeches were delivered from memory, with much vocal inflection, gesturing, eye contact, and animation on the part of the speaker. To ensure that a given speaker lectured in both a dynamic and a static fashion each was coached and evaluated prior to the data collecting by a speech professor at the University. Class responses to a dynamism scale (Andersen, 1961) were obtained immediately after the lecture in order to provide further feedback regarding the extent of treatment differences. The dynamism scale used was a semantic differential which measured audience perception of speaker dynamism.

The other manipulated variable dealt with whether or not the class was directed to pay close attention to the lecture because of a grade threat. Any specific speaker gave a dynamic presentation to a class which had received a grade threat and to a class which had not received such a threat. The same speaker also gave a static delivery to a grade-threatened class and to a class which was not grade threatened.

The class instructors introduced the speakers by saying: "This gentleman is going to talk to you for a few minutes." In classes where the grade threat took place the instructors added: "Pay attention to the lecture because I will expect you to know what he has talked about after he has finished."

In addition to the dynamism scale each class was given a 10-question multiple-choice test immediately after the lecture. This test measured immediate recall of what had been said in the lecture and had a Kuder-Richardson reliability of .64 (Guilford, 1956). Also, the recall test was administered to a control group to see if *S*s had much previous knowledge of interaction analysis.

In order to decrease the effects of a possible lack of equivalence among the error variances of the groups on the results of the analysis of variance, the group sizes were made equivalent (20 per group) by decreasing or increasing the group size as necessary after the data were collected. A group size was decreased through the use of random digit tables (Fisher & Yates, 1963) and increased by considering the mean of the group as an observation the necessary number of times. Six of the groups were decreased in size, one group size was increased by one, and one group size was increased by seven. The equivalence of the group sizes also facilitated the making of meaningful planned comparisons (Hays, 1964), which provided a more powerful statistical analysis than conventional *t* tests.

RESULTS

The planned comparison used to test the difference between the dynamic and nondynamic presentations produced an F ratio of 708 which was significant at the .001 level with $df = 1{,}152$. Clearly, the audiences' perceptions of the dynamic and nondynamic presentations were quite different and in the expected direction.

Table 1 presents mean scores and standard deviations on the immediate-recall test for each of the groups. An analysis of variance for the immediate-recall scores produced F ratios which were significant for all of the main effects and interaction effects at the .01 level with $df = 1{,}152$. However, the only independent variable which was strongly associated with the dependent variable was the

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE IMMEDIATE-RECALL TEST

Group	*M*	*SD*
Dynamic		
Grade threat		
Lecturer 1	9.0	.70
Lecturer 2	8.8	.85
No grade threat		
Lecturer 1	8.8	1.12
Lecturer 2	8.8	1.25
Nondynamic		
Grade threat		
Lecturer 1	6.1	1.97
Lecturer 2	7.2	1.33
No grade threat		
Lecturer 1	7.0	1.84
Lecturer 2	7.3	1.74

mode of presentation. Knowledge about this "speaker dynamism" accounted for 36% of the variance, while no other variable accounted for more than 10%.

The scores on the immediate-recall test for the control group had a mean of 3.88 and a standard deviation of 1.54. A comparison of the mean of the scores of *S*s in the control group with the mean of the scores of *S*s who received a nondynamic lecture treatment and no grade threat was made by computing a Student *t* ratio using a pooled estimate of the population variance. The *t* ratio was found to be 14.38 with $df = 58$ (significant at .001 level).

DISCUSSION

Because of a faulty procedure used in one class, results related to the second hypothesis (effect of grade threat) were inconclusive. Also, due to the restricted nature of the sample, the generalizability of results is somewhat limited. However, the data very convincingly supported the major hypothesis. Students did remember much more from the dynamic lecture than from the static one. Such a finding has strong implications for both practice and future research. The work and planning, which are requisite to effective dynamic speaking, are evidently rewarded with a high degree of audience attention. The relation which was demonstrated to exist between speaker dynamism and audience attention might be considered by teachers as well as by paper readers at conventions of professional organizations.

REFERENCES

ANDERSEN, K. E. An experimental study of the interaction of artistic ethos in persuasion. Unpublished doctoral dissertation, University of Wisconsin, 1961.

BILLINGS, M. Duration of attention. *Psychological Review*, 1914, **21**, 124–135.

CHAPMAN, D. W., & BROWN, H. E. The reciprocity between clearness and range of attention. *Journal of General Psychology*, 1935, **13**, 357–366.

FISHER, R. A., & YATES, F. *Statistical tables for biological, agricultural and medical research*. New York: Hafner, 1963.

GUILFORD, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1956.

HAYS, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1964.

PILLSBURY, W. B. Attention. New York: Macmillan, 1908.

(Received April 14, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 4, 192–206

PERCEPTUAL SPAN TRAINING AND READING ACHIEVEMENT OF SCHOOL CHILDREN

BRUCE R. AMBLE[1] AND SIEGMAR MUEHL
Des Moines Child Guidance Center, Iowa
College of Education, State University of Iowa

To determine if a reading program of over 5,000 phrases would help students improve their reading skills and be a useful supplement to regular reading programs, 3 experimental studies were conducted. The phrase-training program was designed to increase perceptual span and help students develop habits of integrative phrase reading. Ss were intermediate grade-school children and remedial readers. Comparisons between experimental and control groups significantly supported the use of the phrase reading program. Both reading comprehension and reading rate improved, depending on the criterion test employed. Reading gains were consistent for low, medium, and superior readers and were maintained in follow-up posttests.

Several investigators (Buswell, 1939; Gilbert, 1953; Taylor, 1958) have provided evidence for increased efficiency of visual and perceptual skills associated with reading during the school years. However, as Taylor has noted, the development of those skills—as measured by various eye movements, span of perception, and rate of comprehension—has been left largely to trial-and-error processes in the school program. Although various controlled reading devices and instrument techniques have been devised to train these skills, opinions regarding their usefulness have been conflicting (Spache, 1958; Taylor, 1958; Tinker, 1946). Research results in this area are also dissonant. Of particular interest in the present study were the research findings related to tachistoscopic training designed to improve reading rate and/or comprehension skills by increasing the accuracy, speed, and span of recognition.

Investigations conducted by Manolakes (1952) and Sutherland (1945) with college and adult readers, and Marvel (1959) with high-school sophomores found tachistoscopic training procedures to improve reading skills were not as effective as other types of reading improvement programs. The studies by Freeburn (1948) with college freshmen, Barmuth and Aker (1961) with sixth-grade students, and Goins (1958) with first-grade elementary school children showed no differences in reading performance for subjects (*S*s) with tachistoscopic training when compared with control *S*s who received either different or no special reading training. Other studies with widely different populations (Dumler, 1958; MacLatchy, 1946; Renshaw, 1945) not using control procedures claimed significant reading gains for tachistoscopically trained *S*s. Weber (1939) found tachistoscopically trained slow-reading college freshmen and a matched group of *S*s trained with Pressey's *Manual of Reading Exercises for Freshmen* showed increases in reading rate when compared to control *S*s on a follow-up test. At the elementary school level, Jones and Van Why (1963) reported favorable reading rate gains after tachistoscopic training for fourth-grade *S*s and favorable reading comprehension gains for fifth-grade *S*s when compared with matched control *S*s who received no special reading training.

---

[1] Now Assistant Professor, Department of Educational Psychology and Guidance, College of Education, Southern Illinois University.

Two studies attempted to assess changes in span of recognition. Sutherland (1945) reported her experimental *S*s made significant gains in reading span, although no tests were made of the control *S*s. Freeburn (1948) found no span differences between experimental and control *S*s after training.

In summary, perceptual span training to improve reading efficiency has been used primarily on older *S*s with highly practiced reading habits or on very young beginning readers. The results of these studies have been equivocal. Methodological problems were evidenced by inadequate control procedures. The amount of tachistoscopic practice with experimental reading materials was not impressively large. In addition, nonsense forms, geometric figures, and digits were often employed as training materials. A review of the published research suggested the need for further investigation using increased amounts of training materials more appropriate to the reading task with less experienced, intermediate grade-school children.

Three experimental studies were conducted to assess the usefulness of tachistoscopically presented phrase-reading materials. In Experiment I matched groups of fifth- and sixth-grade students each had five hours of distributive training. Experiment II was designed to cross-validate the findings of Experiment I. Experiment III investigated the use of phrase-reading training with intermediate grade- and junior high-school students in remedial reading programs.

EXPERIMENT I[2]

The first investigation attempted to provide answers to four basic questions: Does the span of recognition improve with tachistoscopic phrase-reading training? Does reading rate and/or reading comprehension improve with tachistoscopic training? If there is improvement, do low, average, and superior readers improve in the same areas at the same rate? What is the relationship between span of recognition and reading achievement?

Materials

Motion picture films were developed to provide tachistoscopic-type practice phrases and to test phrase-reading skill. The films were projected at sound speed with 16-millimeter projector equipment. Each phrase appeared on the screen for ⅛ second, enough time for a single reading fixation (Tinker, 1929). The letters in the phrases appeared on the screen white against a contrasting grey background, designed for visual comfort under sustained practice conditions. Two sets of phrase-reading films were designed to provide systematic exposure of many phrases.

The first series, designated the 5200 Phrase Reading Film Series[3], consisted of 10 separate training films, with each film containing 260 different phrases. The films were shown twice during training, which provided a total program of 5,200 phrases. In each film the phrases ranged in length from one to five words with a progressive increase of longer phrases from Film 1 to Film 10. There was a 1½-second interval between presentations to allow time for the student to read each phrase and prepare for the next phrase. Phrases appeared on the screen 37 times per minute at this rate.

The second series of phrase films, designated the 420 Phrase Film Series, consisted of 10 training films, with each film containing 21 different phrases. These phrases were selected from every twelfth phrase in the first series. Each film was shown twice, providing a total of 420 practice phrases. There were 20-second intervals between presentations which gave the students time to read each phrase and write it on paper provided for this purpose. The projector was stopped after each phrase appeared on the screen employing the technique suggested by Gilbert (1959a, 1959b). Phrases appeared on the screen three times per minute at this rate.

---

[2] Experiment I was based on a doctoral dissertation (Amble, 1963). The author gratefully acknowledges the assistance of Siegmar Muehl and James B. Stroud, Cochairman. The cooperation of Arnold Hjelle and the Sigourney, Iowa, Public Schools was much appreciated. Experiments II and III were conducted in the Des Moines, Iowa, Public Schools with the support and cooperation of John Harris, Robert Denny, and E. R. Peterson.

[3] Produced by the Audiovisual Center, Division of Extension and University Services, State University of Iowa, Iowa City, Iowa.

The phrases used in both film series were selected from third-grade reading materials. They consisted of meaningful units when more than one word in length. Several examples are given: shop, ribbon, department; just one, too much, winter came; the long night, an old boat, the next morning; say it out loud, the day was warm, under the apple tree; they looked at each other, a streak of red paint, the red and yellow leaves.

Following each practice film in both series, a similarly designed daily phrase test, also on film, was administered to assess the immediate effects of training, and to establish and maintain motivation. These tests each contained 15 phrases, from one to five words in length. During the daily phrase test, the projector was stopped after each phrase appeared on the screen and *S*s were given time to write out their responses.

Three alternate-form criterion phrase tests (designated A, B, and C) were developed to assess changes in phrase-reading performance (span of recognition) before and after training. Each test contained seven single-word warm-up phrases and twenty-five test phrases. The test phrases ranged in length from one to five words. The criterion phrase tests were presented in the same manner as the daily phrase tests. The warm-up phrases were provided to standardize test conditions, and during the pretest were to screen *S*s not able to read single-word stimuli from the screen. Mimeographed answer sheets were provided *S*s to write out their test responses.

The Iowa Silent Reading Test (ISRT), Elementary Edition, Forms AM, BM, and CM were administered before and after training, to assess changes in reading rate and reading comprehension. Subtest 1, Rate and Comprehension, and Subtest 4, Paragraph Comprehension, were used for this purpose. The Iowa Test Basic Skills (ITBS) is an annually administered school test. Vocabulary and Reading Comprehension test scores were obtained from it for the year preceding and the year following the phrase training program.

Procedures

Subjects. A total of 145 *S*s, 76 fifth- and 69 sixth-grade children, participated in the investigation. The fifth graders ranged in age from 9 years 9 months to 11 years 4 months, with a mean age of 10 years 3 months. For the sixth-grade class, the range was 10 years 11 months to 12 years 10 months, with a mean age of 11 years 6 months. In the final analysis, 120 *S*s were included. Of the 25 eliminated, 8 failed to pass the screening test, 10 were lost in the preliminary matching and 7 were absent during the pretesting.

Design. The study was designed to compare the effectiveness of two approaches of phrase training on subsequent reading achievement. One training method used the 5200 Phrase Film Series. The *S*s assigned to this procedure received a phrase training program of 5,200 *continuous* phrases. The second training group used the 420 Phrase Film Series. These *S*s received a phrase-training program of 420 phrases which were written out during the course of training.

The children in two fifth-grade and two sixth-grade classes in the Sigourney, Iowa, Public Schools were pretested in August at the beginning of the school year with a reading achievement test, ISRT-AM, and Phrase Test A. The *S*s in each grade were ranked in order of reading achievement as determined by an average of the ISRT subtests. Three reading levels were constituted by counting off 20 *S*s at a time, beginning at the top of the distribution in each grade. The levels constituted by this method were designated as the high, medium, and low reading groups. *S*s within each level were assigned alternatively to the 5,200-phrase group and to the 420-phrase group. Thus, the design consisted of two training groups matched at three reading levels, with replications at the fifth and sixth grades.

The students in each classroom received phrase training at the same time each day in their respective programs. The distribution of *S*s both by training groups and reading levels within the four classrooms was a chance factor since the *S*s had been assigned to groups and reading levels independent of class membership. One of the authors and an associate served as training instructors for these periods. The training groups in each class were alternated between the instructors so that over all, each instructor trained about half the *S*s in each group. Location of training was altered so that half the *S*s in the 420-phrase group were trained in the classroom and half in the auditorium. A comparable procedure was employed with the 5,200 phrase group. Film projectors and screen were alternated daily. The training period for all *S*s continued through 10 consecutive school days.

Immediately following the phrase-training program, the ISRT-BM and Phrase Test B were administered. ISRT-CM and Phrase Test C were administered in December, three months after training had been completed.

The instructions for both film series were essentially the same. The *S*s were introduced to the phrase-training program as a practice reading program and were encouraged to improve on the daily phrase tests. Each film was shown twice daily in sequence, and *S*s were encouraged to do better the second time through it. As *S*s became familiar with training and test procedures, it was not necessary to continue reading the detailed instructions. The daily training and test sessions were about 25 minutes in length for *S*s in both experimental conditions.

Raw scores from the daily phrase tests were posted before the homeroom period on large portable charts to maintain motivation. If a student improved on his previous day's score, a gold star was fixed next to his score.

Fourteen *S*s included in the study were absent for one testing period. In these cases the mean score for their training group was used to replace the missing individual's score. On the ITBS fewer matched *S* pairs ($N = 72$) were used in the analysis since complete data were not available for all *S*s.

RESULTS

A four-factor analysis-of-variance design was used with the data (Lindquist, 1956). The two between factors common to all analyses consisted of reading ability levels (high, medium, and low) and school grades (fifth and sixth). The within factors common to all analyses consisted of matched groups (5,200-phrase and 420-phrase groups) and test periods (pretest, posttest, and postposttest). Each of the criterion measures was analyzed separately including daily and criterion phrase tests; ISRT—Rate, Comprehension, and Paragraph Comprehension; and ITBS—Vocabulary and Reading Comprehension. Phrase test scores used in the analysis were raw scores. Standard scores were used on the ISRT and grade-equivalent scores were used on the ITBS.

To make for economy in reporting the results of the various analyses, the main effects associated with reading level and test periods can be summarized immediately, since these effects were similar in all analyses. Reading achievement levels and reading performance increases over test periods were significant in the expected direction ($p < .001$).

The main effect of grade placement was significant in the expected direction in two analyses: ISRT Comprehension ($p < .01$) and ITBS Reading Comprehension ($p < .025$). The interactions of test periods with reading levels or grade placement will not be reported since these scores are collapsed across training groups, and for this reason are not of interest.

Daily Phrase Tests

The analysis showed a significant triple interaction associated with training groups, grades, and test periods ($p < .05$). The interaction indicated that the sixth-grade *S*s in the 5,200-phrase group made more continuous gains during the training program than did the other *S*s.

Criterion Phrase Tests

The analysis showed a significant triple interaction associated with training groups, reading levels, and test periods ($p < .05$). The character of this interaction is shown in Table 1. The *t* tests indicated that the medium reading level, 5,200-phrase group, read significantly more of the phrases immediately following training than did the medium level, 420-phrase group.

TABLE 1

PERFORMANCE OF 5,200- AND 420-PHRASE GROUPS ON CRITERION PHRASE TESTS (EXPERIMENT I)

Training groups	*N*	Reading level	Aug. pretest	Sept. posttest	Dec. post-posttest
5,200 phrase	20	Low	32.23[a]	37.70	40.50
	20	Medium	42.90	51.53	51.30
	20	High	48.93	54.98	55.30
420 phrase	20	Low	27.78	36.88	39.35
	20	Medium	39.88	41.10	47.55
	20	High	52.75	57.70	56.28

[a] Maximum score: 75 words.

However, they did not maintain the advantage during the follow-up posttesting.

In light of the fact that levels were based on reading achievement, it is of interest to note the positive relation between individual phrase test performances and reading levels. Correlations between criterion phrase test scores and the ISRT subtest administered at the same testing period ranged from .39 to .62, for all *S*s.

ISRT Rate

There were no reliable differences associated with training groups, nor were there significant interactions with the other variables.

ISRT Paragraph Comprehension

The analysis showed a significant interaction associated with the training groups and test periods ($p < .01$). The character of this interaction is shown in Table 2. Examination of the data indicated that the 5,200-phrase trained *S*s made increasing gains in comprehension scores over the posttest periods. The *t* tests showed the differences between training groups to be reliable on the three-month follow-up posttest ($p < .05$).

TABLE 2

PERFORMANCE OF 5,200- AND 420-PHRASE GROUPS ON ISRT PARAGRAPH COMPREHENSION SUBTEST (EXPERIMENT I)

Training groups	Aug. pretest	Sept. posttest	Dec. post-posttest
5,200 phrase			
M	132.32	143.65	146.22
GE	4.73	5.93	6.26
420 phrase			
M	134.42	140.25	140.10
GE	4.94	5.53	5.51

Note.—$N = 60$ per Training group.

ITBS Vocabulary

There were no reliable differences associated with training groups nor significant interactions among the variables.

ITBS Reading

The analysis showed a significant main effect for training groups ($p < .025$) and a significant Group × Test Period interaction ($p < .025$). The character of these effects is shown in Table 3. Although *S*s in both training groups made significant improvement from one year to the next, the rate of gain for the 5,200-phrase groups was greater. The *t* tests between training-group means showed a reliable difference at mid-year 1962–63 ($p < .01$).

TABLE 3

PERFORMANCE OF 5,200- AND 420-PHRASE GROUPS ON ITBS READING SUBTEST (EXPERIMENT I)

Training groups	Midyear 1961–62	Midyear 1962–63
5,200 phrase		
M	4.82	6.09
420 phrase		
M	4.78	5.56

Note.—$N = 36$ per Training group.

EXPERIMENT II

Experiment II was designed to cross-validate the findings of Experiment I, to determine the effectiveness of a 2,600-phrase reading program, and to test the feasibility of adapting the training materials for use by classroom teachers as part of the regular program of reading instruction.

Experiment II differed from Experiment I in several respects. A different type of control group was used to provide a basis of comparison. The two studies differed with respect to the personnel administering the phrase training programs. They differed slightly with respect to the training material. Experiment II was conducted in the spring of the year when the students had nearly completed the regular reading program, while Experiment I had been conducted in the fall.

Materials

The 5,200 Phrase Reading Film Series described in Experiment I was modified after the first study. An analysis of phrase-reading performance in the initial study indicated that the one- and five-word phrases were not discriminating among intermediate grade-school children, being either too easy or too difficult to read. These phrase lengths were eliminated in revising the film series, leaving only two-, three-, and four-word phrases as part of the training program. The daily phrase tests were shortened to seven phrases, three and four words in length, and used mainly for purposes of motivation. In other ways the basic features of the series remained the same.

The criterion phrase tests (designated A, B, and C) were made more discriminating by eliminating one- and two-word phrases. The revised tests included twenty-one phrases, three, four, and five words in length. The five-word phrases were retained to provide sufficient top for the tests. No changes were made in the seven-word warm-up procedure used to standardize pretesting conditions.

The Iowa Silent Reading Test (ISRT), Elementary Edition, Forms AM, BM, and CM were again used as criterion reading tests. Subtests 1, Rate and Comprehension; 2, Directed Reading; 3, Word Meaning; and 4, Paragraph Comprehension were administered according to the standard instructions and procedures.

Procedures

Subjects. A total of 436 fifth-grade children participated in the investigation. *S*s ranged in age from 10 years 3 months to 11 years 9 months, with a mean age of 10 years 10 months. The *S*s attended five public elementary schools from various social economic areas in Des Moines, Iowa.

Design. The study was designed to assess the effects of two levels of phrase reading training as compared to a control condition. Two training groups and the control group were constituted as follows. Each of the five schools contained three fifth-grade classes. The classes in each school were randomly assigned to one of three groups. Thus, the two training groups and the control were each made up of five different classes (and teachers) representing five different elementary schools. The training groups were designated according to the amount of phrase reading practice they received: 5,200-phrase group and 2,600-phrase group. The control no-phrase group received no phrase-reading training, and the *S*s remained in the usual school program. The *S*s were distributed among the groups as follows: 5,200-phrase group, 144; 2,600-phrase group, 140; and the no-phrase group, 152.

To control for differences in ability level, *S*s within each of the training groups were assigned to one of three subgroups based on their prephrase test scores.[4] Score ranges were selected that would keep at a minimum the number of *S*s dropped to obtain the proportionality necessary for the statistical analysis. Forty *S*s were included in each of the three low-ability subgroups, with scores ranging from 11 to 33 points. Fifty-three *S*s were included in each of the three average-ability subgroups, with scores ranging from 34 to 53 points. Thirty-two *S*s were assigned to each of the three high-ability subgroups, with scores of 54 to 83 points. The training and control groups each contained 125 *S*s with 375 in all groups.

Sixty-one *S*s were eliminated in obtaining proportionate numbers in the subgroups.

---

[4] Ability grouping based on prephrase test performance was used in all comparisons for two reasons. First, this method allowed for comparison of similar groups; second, correlation coefficients computed between phrase test scores and the various ISRT subtest scores, using all *S*s, indicated that ability differences on the phrase test were related to performance on these subtests (r's ranged from .43 to .69).

The majority of *S*s eliminated had missed missed one or more reading tests; and within this group, *S*s were eliminated on a chance basis. Of the 375 *S*s finally included in the study, 32 had been absent for one testing period. In these cases, the mean score for their training group on a particular test was used to replace the missing individual score.

Classroom teachers administered all aspects of the testing and phrase-training program in this investigation. The pretests were administered on March 17 and 18, 1964, including the ISRT-AM and Phrase Test A. The 5,200-phrase group began training on March 19 for ten consecutive school days. *S*s in this group saw each of the ten films in sequence, with each film shown twice on the day of practice. The 2,600-phrase group continued in the regular classroom program for five days and had five days of phrase training beginning on March 26. They saw Films 1, 3, 5, 8, and 10 each twice during this time. The teachers in both training groups were instructed to continue their usual classroom programs except for the phrase-reading training itself. The control group remained in the usual classroom program throughout the ten-day period and received no phrase training. Posttests were administered after training on April 6 and 7, including the ISRT-BM and Phrase Test B. Postposttests were administered six weeks later, one week before the end of the school term, on May 25 and 26, including the ISRT-CM and Phrase Test C.

RESULTS

A three-factor analysis of variance design was used with the data. The two between factors common to all analyses consisted of training groups (5,200-phrase, 2,600-phrase, and no-phrase groups) and ability levels (low, average, and high) as constituted on the prephrase reading test. The within-factor common to all analyses consisted of test periods (pretest, posttest, and postposttest). Separate analyses were made on each criterion measure, including the criterion phrase tests and the standard subtest scores from the ISRT: Rate, Comprehension, Directed Reading, Paragraph Comprehension, and Word Meaning. To obtain additional information on changes in reading rate, analyses were also made on the total number of items attempted on the ISRT Subtests: Comprehension, Directed Reading, Paragraph Comprehension, and Word Meaning.

To make for economy in reporting the results of the various analyses, the main effects associated with ability groups and test periods can be summarized immediately, since these effects were similar in all analyses. Ability level was significant in the expected direction ($p < .001$) as was performance over test periods ($p < .001$). Test period and ability group interactions will not be reported since the scores in these interactions are collapsed across training groups, and, for this reason, are not of interest.

Criterion Phrase Tests

The analysis of variance showed significant effects associated with training groups ($p < .001$), and the Test Period × Training Group interaction ($p < .001$). The character of this interaction is shown in Figure 1. A simple analysis of variance was carried out at each test period (Lindquist, 1956, p. 47). The results of these analyses showed no reliable differences on the pretest among the groups, and significant differences among the groups on the posttest and the postposttest. The t tests between the means indicated that on the posttest the 5,200-phrase group's performance was better than both the other groups ($p < .02$), and the 2,600-phrase group's performance was better than the no-phrase group ($p < .01$). On the postposttest, the 5,200-phrase group was again better than the other two groups ($p < .01$). However, the difference in performance between the 2,600-phrase group and the no-phrase group was not significant.

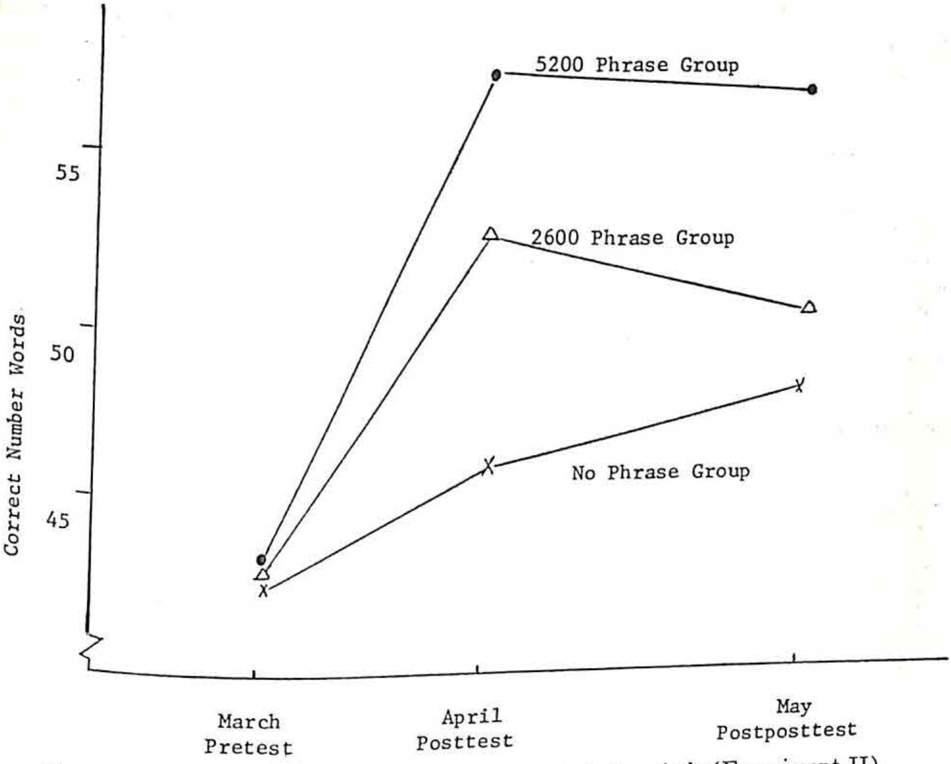


Fig. 1. Criterion phrase test for training groups at test periods (Experiment II).

ISRT Rate

The analysis showed a significant effect associated with training groups ($p < .05$), and the Test Period × Training Group interaction ($p < .001$). Inspection of the means indicated a relative performance increase favoring the 5,200-phrase group. A simple analysis of variance among groups at each test period showed a significant difference on only the postposttest favoring the 5,200-phrase group ($p < .001$).

The characteristics of the interactions in Experiment II are shown in Table 4.

ISRT Comprehension

The analysis showed a significant effect associated with the Test Period × Training Group interaction ($p < .005$). Examination of the mean scores indicated that the interaction resulted from the 5,200-phrase group's marked gain and loss in performance over the test periods.

ISRT Directed Reading

The analysis showed significant effects associated with the Test Period × Training Group interaction ($p < .001$). Examination of the mean scores indicated that the 5,200-phrase group made greater gains over the test periods than did the other two groups. An analysis of variance at each test period showed no reliable differences among the group means on the pretest, a significant difference on the posttest ($p < .005$), and on the postposttest ($p < .005$). On the posttest, t tests between the means showed a significant difference only between the 5,200-

TABLE 4
PERFORMANCE OF 5,200-, 2,600-, AND NO-PHRASE GROUPS ON ISRT SUBTESTS (EXPERIMENT II)

Training groups	March pretest	April posttest	May post-posttest
Reading rate subtest			
5,200 phrase			
M	147.49	161.83	162.11
SD	28.16	25.64	24.91
GE	6.11	9.66	9.73
2,600 phrase			
M	147.86	156.22	150.08
SD	22.64	23.33	21.70
GE	6.34	8.16	6.81
No-phrase			
M	147.78	156.02	151.18
SD	22.54	20.08	19.28
GE	6.33	8.11	7.02
Comprehension subtest			
5,200 phrase			
M	140.65	151.97	146.93
SD	18.76	14.96	16.84
GE	5.53	7.00	6.37
2,600 phrase			
M	142.61	146.68	147.34
SD	17.46	20.41	21.66
GE	5.81	6.34	6.42
No-phrase			
M	143.72	149.89	150.95
SD	16.59	17.54	19.85
GE	5.96	6.74	6.87
Directed reading subtest			
5,200 phrase			
M	144.13	159.93	157.86
SD	18.02	19.85	18.24
GE	6.11	8.79	8.03
2,600 phrase			
M	148.37	152.07	150.57
SD	20.98	19.78	20.10
GE	6.64	7.11	6.91
No-phrase			
M	146.66	156.20	152.39
SD	16.02	16.75	15.83
GE	6.40	7.74	7.14
Paragraph comprehension subtest			
5,200 phrase			
M	145.82	153.14	153.70
SD	21.19	19.81	20.15
GE	6.20	7.32	7.41
2,600 phrase			
M	143.15	148.90	149.17
SD	20.01	19.24	21.61
GE	5.87	6.59	6.72
No-phrase			
M	149.01	151.63	149.98
SD	16.59	16.51	17.99
GE	6.70	7.07	6.80

Note.—$N = 125$ per Training group.

phrase group and 2,600-phrase group ($p < .01$). On the postposttest, *t* tests showed significant differences favoring the 5,200-phrase group when compared to both the 2,600-phrase group and no-phrase group ($p < .02$). The difference between the latter groups was not reliable.

The total number of items attempted on the subtest by *S*s in the three groups was tabulated to determine whether the 5,200-phrase group's gain resulted from increased rate and/or comprehension. The analysis of this data showed a significant Test Period × Training Group interaction ($p < .001$). Inspection of the means indicated that the 5,200-phrase group's gain in correct responses from the pretest to the postposttest (Table 4) was accompanied by an increase in the number of items attempted. During this time period the 5,200-phrase group attempted about 15% more items while making about 25% more correct responses when compared to the other groups.

ISRT Paragraph Comprehension

The analysis showed significant effects associated with training groups ($p < .05$), and Test Period × Training Group interaction ($p < .01$). Inspection of the means indicated that the 5,200-phrase group and the 2,600-phrase group made greater gains when compared with the no-phrase group.[5] The *t* tests between the means for each group from pretest to postposttest indicated that both phrase-trained groups made significant reading gains ($p < .01$), while the no-phrase group made no such gains. Amount of gain appeared to be associated with the extent of phrase training.

The total number of items attempted on the subtest by *S*s in the three groups was tabulated to determine whether these gains in reading comprehension resulted from an increase in rate and/or comprehension. The analysis of these data showed the groups made comparable gains on the number of items attempted over test periods ($p < .001$), but there were no reliable differences among the groups at any of the test periods.

---

[5] Comparisons between training groups were not meaningful on this subtest because of marked differences on the pretest.

ISRT Word Meaning

The reading vocabulary test was used to determine the comparability of intelligence among the training groups. The reasoning in this instance was that performance on a vocabulary test has been shown to be associated with overall intellectual functioning. The pretest scores and the main effects for word meaning favored the 2,600-phrase group and no-phrase group when compared to the 5,200-phrase group ($p < .005$).

EXPERIMENT III

Experiment III was designed to determine if phrase-reading training would be useful with remedial readers, by comparing the phrase-reading program with other remedial reading approaches.

The investigation differed from the two earlier studies in several aspects. The *S*s were children with reading difficulties rather than children in the regular classroom program. The *S*s were older in grade placement and on an average of 2½ years retarded in reading achievement. Both parochial and public school students were included in the current study, and they came from the junior high and high school level as well as the elementary intermediate grades. It was not possible to follow up the investigation with further tests as was done in the earlier studies because the *S*s were from a heterogeneous group of schools and grade levels. The summer reading program was not supported by the ordinary reading requirements available in the regular school program.

Materials

The phrase-reading program was presented with 16-millimeter projection equipment and the phrase-reading films described in Experiment II. The films were projected at silent speed (16 frames per second) for the present study, rather than at sound speed as in the earlier studies. In this manner children with reading difficulties would have more time (2¼″) between phrase presentations, with slightly longer to respond to the phrase as it appeared on the screen (3/16″).[6]

Students in the no-phrase remedial reading program reviewed structure analysis, base words, prefixes, suffixes, compound words, syllabication, dictional skills, vocabulary and compound words, and phonetic and content skills. Oral and silent reading were required throughout the program, with reading comprehension and speed reading exercises. Reading materials and workbooks were provided for students at their level of achievement. The materials used included the *Reading Laboratory Kits* by Science Research Associates; *How to Become a Better Reader* by Paul Witty; *Reading Workbook* by Charles Merrill; and *Phonetics Workbooks* by Lyons and Carnahan. Supplementary readings including *Pleasant Lands, Across the Blue Ridge, Engine Whistles, Trails to Treasure, Parades,* etc.

Procedures

Subjects. The *S*s were public- and parochial-school students attending the summer remedial reading program offered by the Des Moines Public Schools. The program was open to all students in Polk County and an enrollment fee of $8.00 was assessed to defray part of the costs. Attendance was on a voluntary basis.

Grade placement of the students in the study ranged from sixth grade through tenth grade, except for one eleventh grader and one twelfth grader. The average grade placement was 8.3. Fifty-two percent of the students were reading 2½ years or more below grade placement, with an average reading level of 5.8.[7] In addition, 19% of these students were known to have repeated at least one grade in school. The reading level of *S*s ranged in grade placement from 1.9 to 11.2. Relative reading disability was determined by comparing each *S*'s grade placement in school with his level of reading achievement. In this comparison *S*s ranged in relative reading ability from over seven years retarded to three years above expectancy. Several students in the program had certain reading difficulties even though they were able to read at grade placement or above.

---

[6] This procedural change was made because the author believed that children with reading problems might have more difficulty making as rapid a discriminating reading response as students in the regular reading programs. Under these practice conditions *S*s had more time to perceive the phrase on the screen and the next phrase would less likely interfere in the perceptual process.

[7] The school administered the Gates Reading Survey on the first day of summer school. Reading achievement level was determined for each *S* by combining the three subtests (Speed and Accuracy, Reading Vo-

Design. Two experimental training groups were established to compare *S*s in the phrase-reading training program with *S*s in the usual remedial reading program. Three schools were included from different social-economic areas. At each school one instructor taught four reading classes. Two training groups were established within each school by random assignment, so that each instructor had two classes in each group. In this manner each training group was made up of six different classes representing three different school areas, and each instructor taught students in each method of training.

Students who remained in the no-phrase remedial reading program had instruction over a wide range of materials and concepts, and used several established reading approaches. The major goal in this program was to provide meaningful reading experiences for each individual student.

Students assigned to the remedial phrase program had phrase-reading training as part of their daily reading program. The teachers were instructed to show the ten phrase-reading films at least twice during the course of the experiment to establish a minimum training program of 5,200 phrases. Under these instructions, teachers could use phrase training in addition to other remedial reading procedures.

The summer reading program covered 33 daily class meetings over a seven week period. Each class meeting was 64 minutes long and *S*s in both training conditions spent the same amount of time in their respective programs.

All *S*s were pretested with the Iowa Silent Reading Test (ISRT) Form AM, Elementary Edition (Subtests 2–Directed Reading, and 4–Paragraph Comprehension), at the end of the first two weeks of the regular remedial reading program. During the remainder of the program, half of the *S*s remained in the regular remedial program while the other half had the phrase training as part of their program. ISRT-BM was administered at the end of the experiment. Conditions of posttesting paralleled those of pretesting and were similar for *S*s in both training conditions.

Two of the teachers showed each film twice to the phrase classes for a training program of 5,200 phrases. This procedure was comparable to the previous studies. However, the third teacher decided to use the phrase training program almost exclusively. As a result he showed the film series about three and one-half times with his phrase classes, for a training program of 9,800 phrases. For purposes of the statistical analysis, the phrase-reading classes were pooled and the average amount of phrase-reading experiences of 6,500 phrases was used to designate *S*s in this training condition.

The *S*s were distributed in the training groups as follows: the remedial phrase-reading group, 67; and the no-phrase remedial reading group, 61. One hundred twenty *S*s were included in this study, with 60 in each training group to establish the proportionality necessary for the statistical purposes. The eight *S*s eliminated in the process were all absent during the posttests. Three *S*s included in the final selection were absent during the pretests and five *S*s were absent during the posttests. In these cases, the mean score for their training group was used to replace the missing individual score on a particular test.

The *S*s within each of the training groups were assigned to one of three reading levels based on their relative reading disability. Three ability levels were constituted by this method and were designated as mild, moderate, and severe with respect to reading difficulty.

In comparing the two experimental groups, it was noted that the resulting chance assignment of classes to training method did not come out exactly parallel. The remedial phrase-reading group had an average grade placement of 8.7 with an average reading disability of 3.1 years below grade expectation. The no-phrase remedial reading group had an average grade placement of 8.0, with an average reading disability 1.8 below grade expectation. Thus, the phrase group was composed of slightly more advanced students in terms of school grade placement, among which were the more seriously retarded readers.

---

cabulary, and Level of Comprehension) in the manner suggested in the manual.

RESULTS

A three-factor analysis of variance design (Lindquist, 1956) was used on four separate sets of criterion measures, including the standard scores and the total number of responses made on the ISRT Directed Reading and Paragraph Comprehension subtests. The two between factors common to all analyses consisted of training groups (6,500-phrase group and no-phrase group) and ability levels (mild, moderate, and severe with respect to reading disability). The within factor common to all analyses consisted of test periods (pretest and posttest).

Several of the findings can be summarized immediately. The main effects for reading levels were significant on all four criterion measures ($p < .001$), and in each case the *S*s with the least amount of reading disability achieved the highest reading scores. The main effects for test periods showed significant increases on the two Directed Reading analyses (standard scores, $p < .005$; items attempted, $p < .001$), but not on the Paragraph Comprehension analyses. There were significant main effects for training groups favoring the 6,500 phrase groups on the two Directed Reading analyses (standard scores, $p < .005$; items attempted, $p < .01$). The interactions of Test Periods × Reading Levels will not be reported since these scores are collapsed across training groups, and, for this reason, are not of interest.

ISRT Directed Reading

The analysis of the standard scores showed a significant Training Group × Test Period interaction ($p < .05$). The character of this interaction is shown in Table 5. Examination of the means indicated that the interaction resulted from the 6,500-phrase group's marked reading gain when compared with *S*s who received the remedial reading program without supplementary phrase reading training.

TABLE 5

PERFORMANCE OF 6,500- AND NO-PHRASE GROUPS ON ISRT DIRECTED READING SUBTEST (EXPERIMENT III)

Training groups	Pretest	Posttest
6,500 phrase		
M	149.57	155.92
GE	6.76	7.69
No-phrase		
M	144.15	145.02
GE	6.12	6.20

Note.—$N = 60$ per Training group.

The total number of test items attempted were tabulated for all *S*s to determine whether the 6,500-phrase group's gain resulted from an increase in rate and/or comprehension. The analysis showed a trend associated with the Training Group × Test Period interaction. The phrase group made about 12.5% more correct responses from pretest to posttest than did the no-phrase group, while at the same time attempting only 7.5% more items. (All *S*s made correct responses on about 70% of the items attempted.)

ISRT Paragraph Comprehension

The analysis of the standard scores showed a trend associated with the Training Group × Test Period interaction, similar to that found on the Directed Reading subtest. However, the differences between the groups did not reach significance. There were no accompanying changes in reading rate as estimated by the number of items attempted.[8]

---

[8] Standard score means: 6,500 phrase group, pretest 142.42, posttest 146.98; no-phrase group, 143.83 and 144.16, respectively; ($p < .10$). The trend favoring the phrase-reading-trained *S*s on Paragraph Comprehension was in the same direction as the

DISCUSSION

In Experiment I, the tachistoscopically presented, meaningful word phrases helped intermediate grade-school children improve their reading comprehension. After the experimental training period, students assigned to the 5,200-phrase training program were reading significantly better than their matched counterparts in the 420-phrase training program. The magnitude of this difference appears to be sufficiently large to be considered of practical importance. The gains were confirmed not only for the ISRT Paragraph Comprehension Subtest 3 months after training, but comparable findings were evidenced on the ITBS Reading Test.[9]

To determine whether the increased comprehension might be an artifact due to an increase in rate, a tabulation was made of all items attempted on the ISRT Paragraph Comprehension Subtest. From pretest to postposttest, the 5,200-phrase group attempted only 4% more test items while making 13% more correct responses when compared to the 420-phrase group. Furthermore, no reliable differences were found between training groups on the ISRT Rate Subtest or criterion phrase tests. The percentage of students who attempted all test items and therefore might have had an opportunity to go back and recheck some of their responses was almost identical for both experimental groups, about 7%. Therefore, this would not likely account for the differences in reading comprehension. These results suggest that there was an actual gain in reading comprehension independent of changes in reading rate.

Experiment II was designed to cross-validate the findings of Experiment I. In Experiment II, the 5,200-phrase group did significantly better on both post phrase-reading tests than either the 2,600-phrase group or no-phrase group. The changes in phrase reading for *S*s in the 5,200-phrase training program were favorably associated with gains in reading rate and reading comprehension as indicated by the reading achievement test results. The findings were less consistent for students in the 2,600-phrase training program. The difference in the results between the two experiments may be due in part to the changes made on the criterion phrase tests and in part to the differential effectiveness of the revised 5200 Phrase Film Series used in the second investigation.[10]

In Experiment II the 5,200-phrase group made more than 2½ years greater gain in reading on the Rate Subtest when compared to other groups. The reading rate increases were established without loss of reading comprehension, for there were no differences among groups on the Comprehension Subtest over the same materials. The 5,200 - phrase - trained group made over a year greater gain in reading on the Directed Reading Subtest when compared to the other training groups. Differences between the 2,600-phrase group and no-phrase group were not reliable. The reading gains for the 5,200-phrase group were accompanied by significant gains on the number of items attempted when

---

significant findings reported in the earlier two studies.

[9] It appeared unlikely that intellectual differences between groups would be a factor, since *S*s were randomly assigned to training programs at the beginning of the study. To estimate intelligence the two were compared on the ITBS Vocabulary Test. No differences were found.

[10] Since the 5,200-phrase group had a vocabulary disadvantage on the word meaning test, their superior reading gains would not likely be due to some chance assignment of brighter students to their group.

compared to the other groups, indicating differences in the reading scores reflected differential changes in reading rate.

The 5,200-phrase group also made more than a year gain in reading in the Paragraph Comprehension Subtest, while students in the regular school program made no significant improvement. Reading gains for the phrase-trained groups were favorably associated with the amount of practice as indicated by a comparison of the 5,200-phrase group with the 2,600-phrase group. There were no reliable differences among training groups in the total number of items attempted on the Paragraph Comprehension Subtest. Apparently, the greater gains for the phrase-trained groups were due to increases in reading comprehension and not to changes in reading rate. The findings of the 5,200-phrase group were comparable to the findings of the first investigation. In both Experiments I and II the reading progress of the 5,200-phrase group was consistent for low, average, and superior readers and was maintained during follow-up test procedures.

Experiment III was conducted to investigate the use of phrase-reading training with students in a summer remedial reading program. The findings suggest that phrase-reading instruction was a useful supplementary method of remedial reading education, for students in this program showed greater gains on the criterion reading measures. They made over ¾ of a year more gain on the Directed Reading subtest than did the students taught with the usual remedial procedures. There was also a trend on the Paragraph Comprehension subtest favoring the *Ss* who received remedial phrase-reading training. These findings were consistent with the two previous studies. This study was a somewhat more rigorous test of the usefulness of phrase-reading training, however, since the phrase-reading program was compared directly with already established reading approaches.

The Phrase Reading Film Series—Intermediate Grades was developed to provide extensive practice in the rapid reading of meaningful phrases. The phrase-training program was designed to increase the accuracy, speed, and span of recognition of reading materials. The program provided a highly structured learning situation which reduced distracting stimuli during training and provided immediate information on phrase-reading progress. Motivation for reading was increased in the novel learning situation. The results of the research indicated that this technique of phrase-reading practice did significantly enhance reading performance for intermediate-grade children and remedial readers at all reading levels. The reading gains were maintained in follow-up testing procedures after the phrase-training program was completed.

REFERENCES

AMBLE, B. R. Perceptual span training and reading achievement: An experimental study with elementary school children. Unpublished doctoral dissertation, State University of Iowa, 1963.

BARMUTH, T. R. & AKER, C. C. Is the tachistoscope a worthwhile teaching tool? *The Reading Teacher*, 1961, **14**, 172–176.

BUSWELL, G. T. Remedial reading at the college and adult level. *Supplementary Educational Monographs*, 1939, No. 50.

DUMLER, M. A study of factors related to gain in the reading rate of college students trained with tachistoscope and accelerator. *Journal of Educational Research*, 1958, **52**, 27–30.

FREEBURN, C. M. The influence of training in perceptual span and perceptual speed upon reading ability. Unpublished doctoral dissertation, State University of Iowa, 1948.

GILBERT, L. C. Functional motor efficiency of the eyes and its relation to reading.

Publications in Education, Berkeley: University of California Press, 1953, **11,** 159–231.

GILBERT, L. C. Saccadic movements as a factor in visual perception in reading. *Journal of Educational Psychology,* 1959, **50,** 15–19. (a)

GILBERT, L. C. Speed of processing visual stimuli and its relation to reading. *Journal of Educational Psychology,* 1959, **50,** 8–14. (b)

GOINS, J. Visual perceptual ability and early reading progress. *Supplementary Educational Monographs,* 1958, No. 87.

JONES, R. L., & VAN WHY, E. Tachistoscopic training in the fourth and fifth grades. *Journal of Developmental Reading,* Spring, 1963.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1956, 293–296.

MACLATCHY, J. Bexley reading study. *Educational Research Bulletin,* 1946, **25,** 141–168.

MANOLAKES, G. The effect of tachistoscopic training in an adult reading program. *Journal of Applied Psychology,* 1952, **36,** 410–412.

MARVEL, J. Acquisition and retention of reading performance of two response dimensions related to "set" and tachistoscopic training. *Journal of Educational Research,* 1959, **52,** 232–238.

RENSHAW, S. The visual perception and reproduction of forms by tachistoscopic methods. *Journal of Psychology,* 1945, **20,** 217–232.

SPACHE, G. D. A rationale for mechanical methods of improving reading. In O. Causey (Ed.) *College and adult reading improvement.,* 1958, 115–132.

SUTHERLAND, A. J. The relationship between perceptual span and rate of reading. Unpublished doctoral dissertation, State University of Iowa, 1945.

TAYLOR, E. A. The fundamental reading skill. *Journal of Developmental Reading,* 1958, **1,** 21–30.

TINKER, M. A. The study of eye movements in reading. *Psychological Bulletin,* 1946, **43,** 93–120.

TINKER, M. A. Visual apprehension and perception in reading. *Psychological Bulletin,* 1929, **26,** 223–240.

WEBER, C. O. The acquisition and retention of reading skills by college freshmen. *American Journal of Psychology,* 1939, **30,** 453–460.

(Received June 1, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 4, 207–214

DESCRIPTION OF JUNIOR COLLEGES

JAMES M. RICHARDS, JR.,[1] LORRAINE M. RAND, AND LEONARD P. RAND
American College Testing Program, Iowa City, Iowa

In a population of 581 accredited junior colleges, measures of 36 major attributes were intercorrelated. With unity in the diagonal, a principal components analysis was carried out, extracting 12 factors with an eigenvalue greater than 1.00. The first 6 of these factors were rotated to a final solution through the Varimax procedure. The 6 rotated factors were titled: Cultural Affluence, Technological Specialization, Size, Age, Transfer Emphasis, and Business Orientation. The junior college factors are not congruent with factors for 4-year colleges.

In the past 7 years, several ways of describing institutions of higher education have been tried. Pace and Stern (1958) have developed the College Characteristics Index (CCI), a true-false inventory which measures 30 features of the environmental "press" of the college. Astin and Holland (1961) have developed the Environmental Assessment Technique (EAT), which attempts to assess the environment in terms of eight characteristics of the student body: its size, average intelligence, and six "personal orientations" —Realistic, Intellectual, Social, Conventional, Enterprising, and Artistic— based on the proportion of students in each of six classes of major field. These EAT variables were found to account for a substantial amount of variance in CCI scales, and later they were shown to predict the "effects" of the college as reported by the student (Astin, 1963). Still another way to describe college environments is factor analysis of various measures of college characteristics (Astin, 1962, 1965b). Finally, college environments have been viewed simply as a set of potential stimuli, or "observable characteristics of the college that are capable of changing the sensory input to the student attending the college" (Astin, 1965a).

Previous studies of college environments, however, have been restricted to four-year colleges granting the baccalaureate degree. The nearly 600 accredited junior colleges in the United States for the most part have been neglected, although there have been general treatments of junior colleges (Blocker, Plummer, & Richardson, 1965; Medsker, 1960), sociological studies of single junior colleges (Clark, 1960), and studies of the articulation between two- and four-year colleges (Knoell & Medsker, 1964).

The basic purpose of the present study is to develop a description of junior college environments by organizing the information currently available into a brief profile. Such a brief profile can be used both to characterize individual junior colleges, and in subsequent research to study the effects of colleges on students more efficiently. The basic technique is a factor analysis of 36 measures of junior college characteristics. This study, therefore, is largely a replication in a population of junior colleges of Astin's (1962) study of four-year colleges.

PROCEDURE

Population of Junior Colleges

The group of junior colleges consisted of 581 accredited, two-year colleges. This group included all junior colleges for which data are reported in *American Junior Colleges* (Gleazer, 1963),[2] with the exception of colleges which are exclusively for the training of priests, members of religious orders, etc. The sole restriction for inclusion in *American Junior Colleges* is that the college be recognized by regional or state accrediting agencies. Therefore, the group of colleges studied should be considered the population of accredited junior colleges, rather than a sample of some population.

---

[1] The authors are indebted to Max R. Raines, Renee M. Huntley, and H. Bradley Sagen for their critical reading of the manuscript.

Measures of Junior College Characteristics

Thirty-six institutional variables were selected for study. The choice of variables had two primary aims: First, to include at least some data for all methods which are currently used in characterizing institutions, and, second, to include as many as possible of the variables Astin (1962) used in his study of four-year colleges (and to use the same scoring system he used). Unless stated otherwise, the information about junior college characteristics was obtained from *American Junior Colleges* (Gleazer, 1963). In most cases the information in this compendium was reported by each junior college for the academic year 1961–62. The number of colleges for which data were available is indicated after each variable.

Type characteristics. Among the most commonly used ways of classifying colleges are type of control and curricular emphasis. The following five measures of these characteristics were included in this study:

1. Private versus Public Control—Public score 0; private score 1. ($N = 581$.)

2. Degree of Religious Control—Nondenominational score 0, Protestant score 1, Catholic score 2. This scoring system is the same as the one Astin (1962) used in his study of four-year colleges. ($N = 581$.)

3. Liberal Arts Emphasis—No liberal arts curriculum 0, liberal arts plus other curricula 1, liberal arts curriculum only 2. ($N = 574$.)

4. Teacher Training Emphasis—No teacher training 0, teacher training plus other training 1, teacher training only 2. ($N = 574$.)

5. Technical Training Emphasis—No technical school training 0, technical training plus other training 1, technical school only 2. ($N = 574$.)

Financial characteristics. Measures of five financial characteristics were included. In order to eliminate any correlations due simply to differing sizes of junior colleges, all financial characteristics (except tuition) were divided by the junior college's total enrollment, thus expressing each measure on a "dollars per student" basis.

6. Tuition—For public institutions, nonresident fees were used. ($N = 533$.)

7. Endowment—Estimated market value. ($N = 509$.)

8. Operating Budget—Annual expenditures for educational and general purposes. ($N = 482$.)

9. Capital Income—Gifts and appropriations for capital purposes. ($N = 191$.)

10. Scholarship Funds—Amount of money available for scholarships. ($N = 520$.)

Student characteristics. The following fourteen characteristics of the student body were assessed as follows:

11. Percentage of Males in the Student Body. ($N = 571$.)

12. Percentage of Out-of-State Students in the Student Body. ($N = 501$.)

13. Percentage of Foreign Students in the Student Body. ($N = 578$.)

14. Percentage of Part-Time Students in the Student Body. ($N = 571$.)

15. Percentage of Students Earning Half or more of their College Expenses. ($N = 386$.)

16. Total Enrollment—In order to obtain a more nearly normal distribution, the score on this variable is the square root of the total number of students enrolled. ($N = 579$.)

17. Aptitude Level—The score used for this variable was the average composite score on the American College Testing Program's national test battery of applicants to each college in the academic year 1962–63. Unpublished ACT research indicates a correlation of .96 between average composite scores of applicants and average composite scores of freshmen who actually enter colleges. The ACT test battery is a typical test of academic potential, with reliabilities and validities against grade criteria of the magnitude to be expected for such tests (*ACT Technical Report*, 1965). ($N = 314$.)

18. Realistic Orientation—Percentage of students studying agriculture, forestry, engineering, etc. ($N = 497$.)

19. Intellectual Orientation—Percentage

---

[2] It should be noted that more than 200 junior colleges have been established since the publishing of this compendium. At the present time—1966—no comprehensive source of information about these colleges is available.

of students studying science, mathematics, philosophy, etc. ($N = 497$.)

20. Social Orientation—Percentage studying education, nursing, etc. ($N = 497$.)

21. Conventional Orientation—Percentage studying accounting, secretarial, etc. ($N = 496$.)

22. Enterprising Orientation—Percentage studying political science, pre-law, business administration, marketing, etc. ($N = 497$.)

23. Artistic Orientation—Percentage studying art, music, journalism, etc. ($N = 496$.)

Variables 18–23 compose the heart of the Environmental Assessment Technique (Astin & Holland, 1961). There is some doubt as to the appropriateness of using these variables in a factor analysis, since there is an ipsative relationship among them. Nevertheless, they were included in this study because they tap important information and in order to replicate as nearly as possible Astin's (1962) study of four-year colleges. Astin's results, in addition, were clear and meaningful, which suggests that the results were not seriously affected by the ipsative scoring of the EAT variables. In the present study, the EAT variables are based only on fields which clearly belonged in one of the types. Students in an undifferentiated "liberal arts" curriculum were not considered.

24. Homogeneity—Score on this variable is the difference between the highest and lowest EAT variable. High scoring (homogeneous) colleges tend to have students in only one curriculum, while low scoring (heterogeneous) colleges have students enrolled in a wide variety of fields. ($N = 499$.)

Faculty characteristics. These included:

25. Percentage of Faculty Holding a Doctoral Degree. ($N = 575$.)

26. Percentage of Faculty Holding a Master's Degree. ($N = 577$.)

Variables 25 and 26 concern the extent to which the faculty has training beyond the baccalaureate degree. On an overall basis, about one-third of the members of junior college faculties have only a bachelor's degree. Scores on these two variables differ from Astin's (1962) study of four-year colleges in that, in this study, they were based on the total faculty, both full-time and part-time.

27. Percentage of Faculty which is Full-Time. ($N = 579$.)

28. Faculty-Student Ratio—Number of full-time faculty divided by number of full-time students. ($N = 575$.)

Miscellaneous characteristics. Included here are:

29. Library Size—Number of books in the library. ($N = 570$.)

30. Relative Library Size—Number of books divided by the total enrollment. ($N = 569$.)

31. Variety of Curriculum—Total number of different fields of study offered. ($N = 575$.)

32. Percentage of Graduates Going On to Four-Year Colleges. ($N = 467$.)

33. Growth Rate—Percentage of increase in enrollment between 1958 and 1962. ($N = 476$.)

34. California Location—Colleges located in California 1, other colleges 0. This variable was included because California has the most extensive junior college system in the country, and we wanted to investigate whether this system has any special characteristics which distinguish it from other junior colleges. ($N = 581$.)

35. Age of Institution—Colleges founded since 1954 scored 0, colleges founded between 1945 and 1954 scored 1, colleges founded between 1930 and 1944 scored 2, and colleges founded before 1930 scored 3. ($N = 581$.)

36. Placement Service—Colleges having a placement service scored 1, other colleges scored 0. ($N = 581$.)

METHOD

Product moment correlations were computed among the 36 variables.[3] Since not all scores were available for all colleges, a program which allows for missing data was used. Thus correlations are based only on those colleges for which data were available. This procedure *may* bias the correlations in unknown ways. For example, it is possible that the 314 colleges for which ACT test scores are available are quite different from the other 267 colleges.

The resultant correlation matrix was factored by the principal components method based on eigenvalues and eigenvectors with unity in the diagonal and extraction of all factors with an eigenvalue greater than 1.00. This procedure, including the use of unity in the diagonal, is Harris's (1964) *Model A* factor analysis and it follows the rationale presented by Kaiser (1960). A major advantage of this procedure is that it produces factors which are linear combinations of the observable variables, thus making it possible to compute factor scores (Kaiser, 1965).

---

[3] All computations for this study were carried out at the University of Utah Computer Center.

Twelve factors with an eigenvalue greater than 1.00 were extracted. However, several considerations—including an unsatisfactory preliminary rotation of all twelve factors, a comparison of communalities after extraction of each factor with the highest correlation for each variable, and a plot of the eigenvalues—suggested that only half of these twelve factors should be included in the factor rotation. Accordingly, the first six factors were rotated to a final solution by the Varimax procedure (Kaiser, 1958).

Results

Some of the variables involved in this study were highly skewed. For the most part, these variables are those on which a relatively large number of colleges had a score of zero. For example, many junior colleges (especially public colleges) have no endowment. As a result the distribution of endowment is quite skewed. In such cases, of course, there is no transformation which will eliminate the skewness. The correlations among the various junior college

TABLE 1

Unrotated Factor Matrix, Eigenvalues, and Communalities

Variable	Factors												h^2
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	
1. Private (versus public) control	76	11	16	09	−02	−06	−19	03	−01	10	−23	−21	78
2. Degree religious control	58	23	01	03	−07	19	−12	02	06	28	−19	−40	72
3. Liberal arts emphasis	16	48	−28	27	−11	−11	06	33	24	−16	09	06	63
4. Teacher training emphasis	−29	58	−16	−27	−22	−06	13	−32	−09	26	10	−07	78
5. Technological emphasis	−65	−12	12	−38	−14	−06	−16	−10	01	01	−08	04	66
6. Tuition	44	−25	09	11	02	−48	−05	−18	−15	03	00	07	56
7. Endowment/student	26	02	23	10	−25	10	−29	−39	22	−35	01	−10	62
8. Operating budget/student	57	01	31	−12	−09	−22	13	03	−19	−30	16	14	68
9. Capital income/student	17	−18	02	−15	−38	04	38	06	−06	−19	47	−26	71
10. Scholarship funds/student	32	09	07	05	−32	19	−30	−25	41	−15	22	06	65
11. Percentage of males	−47	−34	−08	−39	−16	05	−30	28	07	−03	−04	11	70
12. Percentage of out-of-state students	70	20	27	00	01	−24	15	04	−07	−07	−12	02	72
13. Percentage of foreign students	35	15	33	05	−33	24	−31	−07	24	09	02	20	63
14. Percentage of part-time students	−66	−03	14	44	−10	17	09	06	−12	10	−07	−07	74
15. Percentage of students earning ½ of expenses	−41	02	−06	33	−31	12	33	−18	01	−21	−16	22	64
16. Total enrollment	−63	26	50	18	14	−02	−04	14	05	11	−02	−06	80
17. Aptitude level of students	−13	10	30	−05	−37	−30	47	05	31	01	−17	14	71
18. Realistic orientation	−38	−62	24	−40	−17	00	−10	18	09	−14	−05	−06	84
19. Intellectual orientation	−27	32	−04	03	08	16	−26	−13	−29	−24	15	−34	57
20. Social orientation	13	55	−21	−34	−33	00	25	−14	−10	32	−08	−12	80
21. Conventional orientation	29	−01	−07	43	49	28	23	−01	34	08	18	09	80
22. Enterprising orientation	−02	−02	−04	33	−15	−50	−34	05	−14	20	02	26	63
23. Artistic orientation	13	28	22	10	10	00	06	−15	−35	−44	−31	09	67
24. Homogeneity of environment	39	−59	11	−15	04	12	24	08	19	14	−09	−02	68
25. Percentage of faculty with PhD	03	07	14	28	−36	−31	−06	33	00	09	30	−26	61
26. Percentage of faculty with Master's	−08	48	−53	−10	09	−09	−01	−02	20	−13	09	23	66
27. Percentage of faculty which is full-time	23	19	18	−32	34	−18	−18	−01	−01	08	48	13	65
28. Faculty-student ratio	29	−11	20	02	−15	41	−04	13	−47	19	25	34	78
29. Library size	−13	48	54	01	11	−03	−01	28	07	05	01	−02	64
30. Relative library size	53	16	10	−10	−24	44	04	12	−18	16	−09	29	74
31. Variety of curriculum	−60	51	20	−12	05	02	−02	−15	−07	00	07	07	72
32. Percentage of graduates going to 4-year colleges	−06	42	−41	03	−18	07	−18	44	−02	−18	−06	14	67
33. Growth rate	−11	−18	−14	40	−26	−09	−28	−02	−07	13	−17	−04	43
34. California location	−44	26	54	11	05	16	07	18	08	−05	07	−11	64
35. Age of college	23	39	19	−48	20	−06	−12	13	21	−01	−27	06	68
36. Placement service	−37	−07	29	13	02	−12	−01	−36	14	31	14	19	55
Eigenvalue	5.70	3.44	2.39	2.08	1.70	1.56	1.52	1.32	1.29	1.21	1.18	1.07	

TABLE 2

VARIMAX ROTATION OF FIRST SIX FACTORS

Variable	A	B[a]	C	D[a]	E[a]	F[a]	h²
	Cultural affluence	Technological specialization	Size	Age	Transfer emphasis	Business orientation	
1. Private (versus public) control	47	−41	−22	37	−09	23	63
2. Degree of religious control	47	−33	−17	23	11	−03	43
3. Liberal arts emphasis	06	−37	−03	−06	49	21	43
4. Teacher training emphasis	−05	24	22	08	68	00	58
5. Technological emphasis	−28	67	26	−09	02	−09	61
6. Tuition	−01	−17	−29	24	−30	49	50
7. Endowment/student	42	−05	04	00	−08	13	20
8. Operating budget/student	36	−09	−12	45	−17	32	49
9. Capital income/student	32	26	−23	−03	−04	10	23
10. Scholarship funds/student	48	−05	−09	−02	07	05	25
11. Percentage of males	−21	64	−06	−16	−07	−19	51
12. Percentage of out-of-state students	39	−34	−10	52	−06	35	67
13. Percentage of foreign students	64	−06	10	06	−02	07	43
14. Percentage of part-time students	−17	08	50	−64	−05	−02	69
15. Percentage of students earning ½ of expenses	00	09	20	−56	14	06	38
16. Total enrollment	−22	09	83	−13	−04	03	76
17. Aptitude level of students	12	30	23	03	06	42	34
18. Realistic orientation	−11	73	−02	−09	−45	−08	77
19. Intellectual orientation	−09	−05	28	−08	27	−19	21
20. Social orientation	25	15	−06	23	67	02	59
21. Conventional orientation	−02	−68	−03	−05	−22	−26	59
22. Enterprising orientation	−17	−10	−03	−15	02	57	38
23. Artistic orientation	16	−22	31	21	00	08	22
24. Homogeneity of environment	21	08	−40	12	−56	−11	55
25. Percentage of faculty with PhD	14	00	08	−16	05	53	33
26. Percentage of faculty with Master's	−27	−14	−12	07	65	−09	54
27. Percentage of faculty which is full-time	−08	−08	06	60	−02	−02	37
28. Faculty-student ratio	50	−04	−03	−04	−20	−20	33
29. Library size	11	−07	67	27	09	08	56
30. Relative library size	69	−13	−13	15	08	−20	58
31. Variety of curriculum	−21	21	66	01	37	−10	68
32. Percentage of graduates going to 4-year colleges	01	−08	−07	−14	60	−05	39
33. Growth rate	−01	−05	−10	−47	−03	26	30
34. California location	04	08	75	−08	−05	−07	58
35. Age of college	09	03	13	67	20	−11	52
36. Placement service	−16	14	37	−15	−18	13	25

[a] Reflected factor.

TABLE 3

SIMILARITY BETWEEN FACTORS FOR JUNIOR COLLEGES AND FOR FOUR-YEAR COLLEGES

4-year college factors	Junior college factors					
	Cultural affluence	Technological specialization	Size	Age	Transfer emphasis	Business orientation
Affluence	.5987	.0108	.0658	.1968	−.1304	.4580
Realistic orientation	−.1971	.6639	.1265	−.0759	−.5135	−.1808
Size	−.0865	.3287	.7236	.1098	.0815	.0654
Masculinity	−.3888	.3677	.1195	−.3507	−.2957	−.2998
Homogeneity	.1394	.2877	−.5656	.0277	−.4278	−.1736
Public vs. private control	.4410	−.4806	−.2397	.3302	−.1921	.3542

Note.—4-year college factors obtained from study by Astin (1962).

characteristics were computed[4] and subjected to factor analysis. The unrotated matrix of the twelve factors with an eigenvalue greater than 1.00 is presented in Table 1. The factor solution for the Varimax rotation of the first six factors is given in Table 2.

Finally, in order to compare the factors obtained in this study with the factors Astin (1962) obtained for four-year colleges, the *Coefficient of Congruence* (Tucker, 1951) was computed between each rotated factor for this study and each of Astin's rotated factors. These calculations involved only the variables common to the two studies. Results are shown in Table 3, with Astin's factors rearranged so that, to the extent possible, highest *Coefficients of Congruence* are in the diagonal.

DISCUSSION

The rotated factors are briefly described and interpreted below:

Factor A

The variables with high loadings on this factor describe a college which has a large number of library books per student, relatively many foreign and out-of-state students, and many faculty members relative to the number of students. It is privately or religiously controlled, and is relatively well financed. This pattern looks like the factor named Affluence by Astin (1962) in his study of four-year colleges. In the present study, however, the factor appears to involve facilities, such as the library and the faculty, more than financial wealth. An appropriate title, therefore, would be *Cultural Affluence.*

Factor B

Loadings on this factor describe a college with a technological emphasis and many students in technical programs, with many male students, with few students studying such fields as education and secretarial work, and with few out-of-state students. It is a public school which does not emphasize the liberal arts. A good title would be *Technological Specialization.*

Factor C

Colleges that would score high on this factor have large enrollments; a varied, heterogeneous curriculum; many part-time students; and a placement service. The best title for this pattern would probably be *Size.* The college scoring high would probably be an urban-centered, open-door, comprehensive college, with a strong emphasis on continuing education. In addition, one would expect the high scoring college to be characterized by an impersonal atmosphere, few personal contacts between students and faculty, several highly organized student subcultures, and a relatively clear status hierarchy of social groups.

California junior colleges are distinguished from other colleges by a high loading on this factor. It is interesting, and probably contrary to popular belief, that California location failed to load substantially on any other factor.

Factor D

Loadings represent a college which is old, which has faculty and students who are both full-time, which has few working students but relatively many out-of-state students, which has *not* grown, which spends relatively much money per student, and which is a private school. An appropriate title would be *Age*. The high-scoring college would probably resemble a small, four-year liberal arts college. It would likely have many traditions, a residential student body, and an administration which conceived of its role as acting *in loco parentis*. Such a college would also be likely to have a selective admissions policy, although not necessarily one that emphasizes academic aptitude. An alternative title, therefore, might be Traditional Exclusiveness.

---

[4] Tables showing Means, Standard Deviations, and Numbers of Observations for Junior College Characteristics and Correlations among Junior College Characteristics have been deposited with the American Documentation Institute. Order Document No. 8921 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $1.25 for microfilm or $1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

Factor E

Colleges characterized by the variables loading high on this factor emphasize teacher training and liberal arts and offer a heterogeneous environment. They have many students studying such fields as education, many graduates who go on to four-year colleges, and many faculty members with master's degrees. A common denominator to most of these variables is a requirement for further education beyond junior college, and, accordingly, many graduates of high-scoring colleges seek advanced training. The best title for this factor would probably be *Transfer Emphasis*.

One would expect the colleges scoring *low* on this factor to be terminal colleges primarily concerned with practical vocational training, making little effort to model their curriculum on what has been traditional for four-year colleges. The high-scoring college would be concerned more with preprofessional, exploratory training.

Factor F

The high-scoring college on this factor has relatively many students in fields characterized as Enterprising, relatively many faculty members with Ph.D.'s, high tuition, bright students, and many out-of-state students. In addition, it spends an above-average amount of money per student. The interpretation of this factor is less manifest than was the case for the preceding factors. As an aid in the interpretation, colleges with high scores on the Enterprising Orientation variable were identified. Two types of colleges appeared to predominate. First, small private colleges on the East coast with many students studying sales and retailing, and, second, large public colleges on the West coast with many students studying management. The trait common to these two kinds of colleges appears to be an emphasis on providing students with a *business* skill having immediate utilitarian value. This factor, therefore, might best be named *Business Orientation*.

The comparison of the factors for this study with Astin's (1962) factors for four-year colleges indicate some similarity, but in no case was the *Coefficient of Congruence* high enough to justify considering factors identical. In the present study, of course, no attempt was made to use Astin's solution as a criterion in the rotation of the junior college factors. On the other hand, an independent analytic rotation is probably a more stringent test of equivalence of factors than is using one solution as a criterion for the rotation of another solution and independent analytic rotations do produce good matches between rotated factors in some cases (Richards, 1966; in press). Therefore, these results appear to mean that junior colleges are different from four-year colleges, and that it would not be appropriate to apply a classification scheme developed for one type of college to the other type.

The primary goal of this study was to provide a brief profile which can be used to characterize junior colleges, and which will make possible more efficient research on the effects of junior colleges on their students. It seems clear that this goal was attained, for the original 36 scores were reduced to six factors which are reasonably clear and easily interpreted. The reduction to six representative factors provides a simple, economical set of items for assessing junior college environments in research on the ways in which different colleges affect student accomplishment and growth. We hope, however, that this profile will be regarded only as a first step, that junior colleges will now receive the attention from researchers that is warranted by their importance to society, and that this attention will in turn lead to better descriptive schemes, better studies of junior college effects, and, ultimately, to better junior colleges.

REFERENCES

American College Testing Program. *ACT Technical Report.* Iowa City: Author, 1965.

ASTIN, A. W. An empirical characterization of higher educational institutions. *Journal of Educational Psychology*, 1962, **53**, 224–235.

ASTIN, A. W. Further validation of the Environmental Assessment Technique. *Journal of Educational Psychology*, 1963, **54**, 217–226.

ASTIN, A. W. Recent studies of college environments. Paper read at American Personnel and Guidance Association, Minneapolis, April, 1965. (a)

ASTIN, A. W. *Who goes where to college.* Chicago: Science Research Associates, 1965. (b)

ASTIN, A. W., & HOLLAND, J. L. The Environmental Assessment Technique: A way to measure college environments. *Journal of Educational Psychology*, 1961, **52**, 308–316.

BLOCKER, C. E., PLUMMER, R. H., & RICHARDSON, R. C., JR., *The two-year college: A social synthesis*. Englewood Cliffs, N.J.: Prentice-Hall, 1965.

CLARK, B. R. *The open door college.* New York: McGraw-Hill, 1960.

GLEAZER, E. J., JR. (ED.) *American junior colleges.* Washington, D. C.: American Council on Education, 1963.

HARRIS, C. W. Four models for factor analysis. Paper read at American Psychological Association, Los Angeles, September, 1964.

KAISER, H. F. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 1960, **20**, 141–151.

KAISER, H. F. Psychometric approaches to factor analysis. In *Proceedings of the 1964 invitational conference on testing problems*. Princeton, New Jersey: Educational Testing Service, 1965.

KAISER, H. F. The Varimax Criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, **23**, 187–200.

KNOELL, D. M., & MEDSKER, L. L. *Factors affecting performance of transfer students from two- to four-year colleges*. Berkeley, California: Center for the Study of Higher Education, 1964.

MEDSKER, L. L. *The junior college: Progress and prospects*. New York: McGraw-Hill, 1960.

PACE, C. R., & STERN, G. G. An approach to the measurement of psychological characteristics of college environments. *Journal of Educational Psychology*, 1958, **49**, 269–277.

RICHARDS, J. M., JR. Life goals of college freshmen. *Journal of Counseling Psychology*, 1966, **13**, 12–20.

RICHARDS, J. M., JR. A factor analytic study of the self-ratings of college freshmen. *Educational and Psychological Measurement*, in press.

SANFORD, N. (ED.) *The American college.* New York: Wiley, 1962.

TUCKER, L. R. A method of synthesis of factor analysis studies. *Personnel Research Section Report*, No. 984. Washington: Dept. of the Army, 1951.

(Received June 21, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 4, 215-219

COMBINING RESPONSE MODES IN PROGRAMMED INSTRUCTION[1]

JOANNA P. WILLIAMS
University of Pennsylvania

A group of 120 6th-grade pupils worked through a linear program in 1 of 4 versions: standard constructed-response format (A); standard multiple-choice format (B); a combination format in which the constructed-response mode was used on frames that required technical terminology as responses and the multiple-choice mode on items where general vocabulary was required (C); and a combination format in which response-mode was uncorrelated with item-type (D). Groups A and C were significantly better than Groups B and D on a constructed-response posttest and on gain scores. The findings suggest that the use of a varied format may not in itself facilitate performance. Rather, the important variable may be the correlation between training mode and type of item.

An instructional method is evaluated according to two general criteria. First, the effectiveness of the method must be assessed. Does it teach what it is supposed to teach? Or, put another way, does performance on the criterion test reach the desired level? Second, the efficiency of the method must be taken into account, for methods that require unduly large amounts of training are not practical.

Constructed-response (CR) training, in which the student is required to write a response to every frame, has compared favorably in effectiveness with all other modes typically investigated (e.g., Holland, 1960; Krumboltz & Weisman, 1962; Williams, 1963). However, this mode is relatively inefficient: Less is learned in a given amount of time than with any other training mode (Goldbeck & Campbell, 1962; Williams, 1963).

One variable that determines the relative effectiveness of different training modes is the type of item used on the criterion test. CR training, for example, is superior to multiple-choice (MC) training on test items that use novel or technical terminology, but not on items that use general, familiar vocabulary. This has been demonstrated with both new and review material, and in both elementary pupils and college students (Williams, 1963, 1965). As usual, however, CR training required more time, and the overall efficiency of that mode was lower.

These findings suggest one way in which it might be possible to increase the efficiency of a program without diminishing its effectiveness: Use CR training on frames requiring technical terminology and MC training on the other items. How would such a program compare in (*a*) effectiveness and (*b*) efficiency with a program composed entirely of CR frames and one which is entirely MC? The present study compares programs of these three types.

In order to evaluate such a combined-mode program properly, one must take account of possible interfering effects introduced by requiring the concurrent use of two response modes. Since the effects of introducing a mixture of training modes are unknown (Taber, Glaser, & Schaefer, 1965), a fourth program was designed, which also used the same two training modes, but in which there was no correlation between training mode and type of item. That is, frames requiring technical responses and those requiring general responses were assigned randomly to CR presentation or to MC presentation.

---

[1] This research was supported in part by the Cooperative Research Program, United States Office of Education (Project #2100). The author is indebted to Ellen I. Levy for her assistance in the collection and analysis of the data, and to Coronet Instructional Films, Chicago, Illinois, for permission to reproduce a portion of their program "Grouping Animals: What is a Mammal?".

METHOD

Subjects

Subjects (*S*s) were 120 sixth graders who had not been previously exposed in school to the material covered in the program. An additional 14 *S*s were dropped: 11 for failure to complete all training and testing sessions, and 3 to equate pretest scores and *N*s in the experimental groups.

Materials

The first 120 frames of the Coronet Program "Grouping Animals: What is a Mammal?" were reproduced in booklet form. Eighty-eight of these frames, 44 requiring technical terms as the response (e.g., *vertebrae* and *phylum*) and 44 requiring familiar vocabulary, were specified for experimental variation. The remaining 32 frames, which were distributed randomly throughout the booklet, appeared as complete statements, one per page, to be read as regular text material. No overt response was required. These frames were presented in this manner in order to balance the number of technical and general items to be manipulated in the four training groups.

In Group A, all 88 frames consisted of the standard CR format. Each frame appeared on a separate page with the correct response presented alone on a following page. The *S* was required to write out the answer for every frame on a separate answer sheet. In Group B, all 88 frames consisted of the MC format. Under each item, presented in the same way as in Group A, two alternative responses were printed. The *S* was required to underline his choice, and he found the correct response on the following page. In Group C, those 44 frames which required technical terminology as the response were presented in the CR format; the other 44 frames, which required general vocabulary, were presented in the MC mode. In Group D, 44 frames, randomly selected from the 88, were presented in the CR format, and the other 44 appeared as MC items.

All *S*s recorded their responses on a separate answer sheet, which contained blanks for the CR items and the two alternative choices for the MC items. The frames that were presented as complete statements and required no response were indicated by a row of *X*'s by the appropriate number.

An objective test was made up of 16 items that required constructed responses. Half the items required *S* to utilize technical terms introduced in the program (such as *vertebrae* and *phylum*), and the other half were items to which *S* could respond with familiar words in his standard vocabulary. An additional 16 items, which required a choice between two alternative answers, were also given; half demanded technical terms as responses, and half were of the familiar vocabulary type. These were analyzed as two separate tests, but they were presented to the subject as one test, in which items of both kinds—CR and MC—were mixed together.

Items were balanced so that every technical term which was used as a response appeared both on the CR and on the MC tests. In addition, the test was divided into two parts, each of which had to be completed independently. In each half, each response term appeared either as an MC or as a CR item, so that there could be no cueing from one type to another. Half the *S*s were given Part 1 first and then Part 2; for the other *S*s, the order was reversed.

Procedure

The *S*s were randomly assigned to one of the four training conditions. All *S*s were given the objective test as a pretest, and sixteen minutes were allowed (eight minutes for each half). On each of two successive days of training, starting three days after the pretest, *S*s worked through 60 frames of the program. The time taken by each *S* to work through each day's assignment was recorded. On the day following the completion of the program, *S*s were again given the objective test with the same time limits as in the pretest. All *S*s finished the test in the time allowed.

Before the experiment began, *S*s were told that this was an experiment in programmed instruction and that their class had been specially chosen for participation. It was pointed out that their test scores would not count as part of their school grades, but that they should try to do as well as possible.

TABLE 1
MEAN NUMBER CORRECT ON CONSTRUCTED-RESPONSE TEST

Training mode	Pretest				Posttest			
	Technical items	General items	Overall M	SD	Technical items	General items	Overall M	SD
A (All constructed response)	0.73	2.23	2.96	2.12	3.30	4.70	8.00	3.90
B (All multiple choice)	0.60	1.97	2.57	2.03	2.07	4.03	6.10	3.21
C (Combined, correlated)	0.50	2.17	2.67	1.88	2.87	4.13	7.00	3.72
D (Combined, uncorrelated)	0.57	2.10	2.67	1.65	2.40	3.70	6.10	3.05

RESULTS

Means and standard deviations of the performance measures on the CR criterion test are presented in Table 1. Both pretest and posttest scores were normally distributed. There was a significant increase from pretest to posttest (The ts ranged from 4.97 to 6.06, $df = 58$, $p < .01$). An untrained control group was considered unnecessary since previous work (Williams, 1965) indicated no difference between pretest and posttest following similar manipulations.

Table 1 also presents the mean number of correct responses on the CR criterion test as a function of training mode and type of test item. As indicated by analysis of variance, there were no differences among the training mode groups on the pretest ($F < 1$). Scores on general vocabulary items were higher than those for technical items ($F = 105.34$, $df = 1/116$, $p < .001$). There were no significant interactions ($F < 1$). On the posttest, as on the pretest, type of test item (general > technical) showed a significant effect ($F = 84.63$, $df = 1/116$, $p < .001$). In addition, the differences due to training mode reached the .10 level of significance ($F = 2.02$, $df = 3/116$). The two variables did not interact ($F = 1.03$, $df = 3/116$).

Since a priori predictions had been made about the training mode variable, specific comparisons were made. They indicated that Group A (all CR) was not significantly different from Group C (combined modes, correlated with item type) ($F = 1.23$, $df = 1/116$), but that Groups A and C together were significantly higher than Groups B (all MC) and D (combined modes but uncorrelated with item type) ($F = 4.71$, $df = 1/116$, $p < .05$). The latter two groups did not differ ($F < 1$).

The difference between pretest and posttest scores was taken as a measure of the amount gained from the experimental training. Analysis of the gains on the CR test show results very similar to those of the analysis of the CR posttest. Again, training mode reached the .10 level of significance ($F = 2.48$, $df = 3/116$). The two types of item were not significantly different ($F < 1$), and there was no interaction ($F = 1.59$, $df = 3/116$).

Specific comparisons among these gain scores showed the same pattern as did those on the posttest. Group A was not reliably higher than Group C ($F = 1.08$, $df = 1/116$). Groups A and C scored significantly higher than Groups B and D ($F = 6.33$, $df = 1/116$, $p < .05$). The latter two groups did not differ ($F < 1$).

Means and standard deviations of the amount of time taken in training are presented in Table 2. There were no

TABLE 2

TRAINING TIME (IN MIN.)

Training mode	*M*	*SD*
A (All constructed response)	39.4	7.5
B (All multiple choice)	38.6	9.6
C (Combined, correlated)	43.8	10.2
D (Combined, uncorrelated)	40.2	9.2

differences among the experimental groups ($F = 1.82$, $df = 3/116$).

An efficiency score was calculated for each subject by dividing his posttest score by his training time, giving a measure of amount learned per unit of training time (Goldbeck & Campbell, 1962). There were no differences among the groups on this measure ($F = 1.16$, $df = 3/116$). An efficiency measure based on actual gains (difference score divided by training time) also failed to show differences ($F = 1.60$, $df = 3/116$).

There was a significant improvement from pretest to posttest on the MC criterion test ($p < .01$ for each group), but, as in a previous study (Williams, 1965), this test was not sensitive to differences in training mode ($F < 1$, $df = 3/116$, on both posttest and gain scores).

DISCUSSION

The results were as expected in terms of the effectiveness criterion, that is, achievement scores. The superior posttest scores and gain scores of the CR training group (A) as compared with the MC training group (B), previously demonstrated by Fry (1960) and Williams (1963, 1965) on CR tests, also were found in the present study, where straight reading frames were incorporated into the programs. Since this comparison reached significance, the present experimental situation was considered suitable for assessing the relative effectiveness of the combined modes.

The important comparison is that between Groups C and B. As predicted, the program that combined CR training on technical items and MC training on other items (Group C) was more effective than the wholly MC program (Group B), indicating that a combined mode can indeed boost performance. Moreover, it is clear that it was the correlation between CR training and technical terminology that accounted for Group C's superiority, and not the mixed format. Group D, which contained the same mixed format and the same amount of CR training but uncorrelated with item type, was significantly below Group C, and in fact was no better than Group B.

Programmed instruction has been criticized on the grounds that it becomes monotonous and unstimulating (e.g., Fitzgerald, 1962). It is claimed that the routine, highly limited format tends to restrict independent thinking (Doty & Doty, 1964), and such concerns have led to the reasonable suggestion (Taber, Glaser, & Schaefer, 1965) that there be some variation in the format of the program.

The present data suggest, however, that merely introducing a combination of modes into a program on a random basis does not enhance the effectiveness of the program. Unless a mixed-format program is designed to capitalize on the favorable interaction between the specific training mode and the type of item (in this case, CR mode and technical items), then it might be as effective to use a single training mode. Other methods of incorporating variety into a program, however, such as interspersing "test-type" frames that have no formal confirmation (Hershberger's, 1964, self-evaluational response items), might prove useful.

The data are not conclusive with respect to the efficiency criterion. Contrary to expectation, there were no significant differences among the groups in the amount of time taken to go through the program. The addition of the large number of reading frames, which all groups could go through at a rapid pace, probably minimized overall training time differences. Therefore the present conditions did not afford an adequate situation in which to test the hypothesis that a combined mode where mode is correlated with item type would be more efficient than an all-constructed-response program.

REFERENCES

Doty, B. A. & Doty, L. A. Programed instructional effectiveness in relation to certain student characteristics. *Journal of Educational Psychology*, 1964, **55**, 334–338.

Fitzgerald, H. Teaching machines: A demurrer. *School Review*, 1962, **70**, 247–256.

Fry, E. B. A study of teaching machine response modes. In A. A. Lumsdaine & R. Glaser (Eds.), *Teaching machines and programmed learning*. Washington, D. C.: National Education Association, 1960. Pp. 469–474.

Goldbeck, R. A. & Campbell, V. N. The effects of response mode and response difficulty on programed learning. *Journal of Educational Psychology*, 1962, **53**, 110–118.

Hershberger, W. Self-evaluation responding and typographical cueing: Techniques for programing self-instructional reading materials. *Journal of Educational Psychology*, 1964, **55**, 288–296.

Holland, J. G. Design and use of a teaching machine program. Paper read at American Psychological Association, Chicago, September 1960. (Mimeo).

Krumboltz, J. D. & Weisman, R. G. The effect of overt versus covert responding to programed instruction on immediate and delayed retention. *Journal of Educational Psychology*, 1962, **53**, 89–92.

Taber, J. I., Glaser, R. & Schaefer, H. H. *Learning and programmed instruction*. Reading, Mass.: Addison-Wesley Co., 1965.

Williams, J. P. A comparison of several response modes in a review program. *Journal of Educational Psychology*, 1963, **54**, 253–260.

Williams, J. P. Effectiveness of constructed-response and multiple-choice programing modes as a function of test mode. *Journal of Educational Psychology*, 1965, **56**, 111–117.

(Received September 2, 1965).

Journal of Educational Psychology
1966, Vol. 57, No. 4, 220–229

LEARNING AND MEASURED ABILITIES[1]

JAMES P. DUNCANSON
Princeton University

The interrelationships among measures of learning and measures of abilities were investigated by administering a battery of ability tests and learning tasks to 102 6th-grade children. The learning tasks included concept-formation, paired-associate, and rote-memory tasks with verbal, numerical, and figural material. Ability scores and measures of learning were factor analyzed and 7 factors extracted. 1 factor was restricted to the ability measures; 3 were restricted to the learning measures; and 3 were common to both the ability and learning measures. It was concluded that performance in a learning task is related to measured abilities and to performance in other learning tasks and that there are learning factors which are not related to the abilities measured.

The first half of the twentieth century saw considerable interest in the relationship between learning ability and intelligence. Experimenters fairly consistently failed to demonstrate an association between them, however, so that we find statements to the effect that learning ability and intelligence are not the same thing (Guilford, 1940; McGeoch & Irion, 1952; Tyler, 1956; Woodrow, 1946). In addition, learning in one situation appeared to be unrelated to learning in other situations (Guilford, 1940; Husband, 1939, 1941).

These conclusions contradict everyday observation as well as the consistently high relationship found between intelligence and scholastic achievement and the established usefulness of ability testing for selection and prediction. Resolution of this contradiction depends first of all upon adequate measures of learning, and second upon the scope and nature of the learning situations and of the abilities measured.

Measures of learning ability which have been used include final level of attainment, gain (the difference between initial and final levels), and parameters of learning curves. Each of these is subject to more or less serious objections, including the effect of different initial levels, ceiling effects, unequal difficulty at different stages of learning, the choice of the functions to represent the learning process, and the number of parameters required by the function.

Two comparatively recent studies have investigated the relationship between learning and measured abilities using parameters of fitted learning curves. Both Stake (1961) and Allison (1960) found in factor-analytic studies that parameters of learning curves were related to ability and achievement measures and that there were additional factors common to more than one learning situation, but no general factor common to all.

Games (1962) also found that the learning of serial and paired-associate verbal tasks was related to measures of rote and span memory. Additional evidence of a relationship between learning and measured abilities is contributed in the area of motor tasks by Fleishman and Hempel (1954, 1955) and Fleishman (1960).

---

[1] This report is based on a dissertation submitted to Princeton University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The author is indebted to his dissertation committee, particularly Harold Gulliksen, to Miss Ruth E. Parker and other school officials, and to Educational Testing Service. Financial support was received from Office of Naval Research Contract Nonr 1858(15), Harold Gulliksen, Principal Investigator, and Office of Naval Research Contract Nonr 2214(00), John W. French, Principal Investigator.

[2] The author is now at Bell Telephone Laboratories.

It is apparent, then, that abilities do contribute to performance in a learning situation. The principal difference between the present investigation and previous ones is an attempt to provide more adequate measures of learning performance and to systematically vary a small variety of learning tasks. The basic question is the relationship between measures of ability and measures of learning and between measures of learning in one task with those in other tasks.

METHOD

Learning Tasks

Three types of learning tasks and three types of material were combined for a total of nine tasks. The types of task were concept-formation, paired-associates, and rote-memory; the types of material were verbal, numerical, and figural.

The concept-formation tasks were adapted from the Wisconsin Card Sorting task (Berg, 1948). Each task consisted of a series of stimuli printed one to a page, the pages being bound into a booklet. Each stimulus had four attributes which varied in three steps. For example the figural concept-formation task stimuli had one, two, or three borders; one, two, or three figures; the figures were either circles, squares, or triangles; and the figures were either white, shaded, or black. For each task one of the values of one of the attributes was chosen as defining positive instances of the concept. For the example given, the concept was "shadedness"; that is, all stimuli having a shaded figure or figures, whether they were circles, squares, or triangles and regardless of the number of borders, were positive instances of the concept.

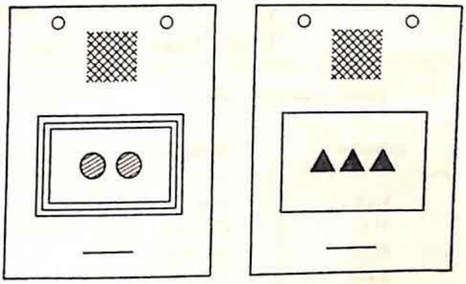


FIG. 1. Positive (left) and negative (right) instances of the concept "shaded" taken from the figural concept-formation task. (The correct response was printed on the back of each page behind the cross-hatched area.)

The attributes and steps of the three concept-formation tasks are presented in Table 1. Figure 1 presents a positive and a negative instance of the concept "shaded." Twenty-five of the possible 27 positive instances and 25 negative instances were chosen and a block of 50 pages prepared. The sequence was random subject to the following restrictions: (*a*) the first stimulus was a positive instance, (*b*) there were five positive and five negative instances in each block of 10, and (*c*) no more than three positive or three negative instances occurred consecutively. The block of 50 stimuli was repeated twice for the figural task and three times for the verbal and numerical tasks to

TABLE 1

ATTRIBUTES AND STEPS USED IN CONSTRUCTING CONCEPT-FORMATION LEARNING TASKS

Concept formation—verbal		Concept formation—numerical		Concept formation—figural	
Attributes	Steps	Attributes	Steps	Attributes	Steps
Number of borders	1, 2, 3	Number of borders	1, 2, 3	Number of borders	1, 2, 3
Number of words	1, 2, 3	Type of border	Solid, sides only, corners only	Number of figures	1, 2, 3
Type of word[a]	Noun, verb[a], adjective	Number of numbers[a]	1, 2[a], 3	Type of figure	Circle, square, triangle
Type of print	All capitals, initial letters capitalized, all lower-case letters	Number of digits per number	1, 2, 3	Brightness of figure[a]	White, shaded[a], black

[a] The particular aspects defining positive instances of each concept.

TABLE 2

PAIRS USED IN PAIRED-ASSOCIATES LEARNING TASKS

Paired associates—verbal[a]		Paired associates—numerical[b]		Paired associates—figural	
Stimulus	Response	Stimulus	Response	Stimulus	Response
YAT	jewel	POLEF	63	1	□
TIS	dinner	GOKEM	95	2	ω
ZUG	money	LATUK	57	3	+
KEM	village	ZUMAP	89	4	>
SOZ	insect	TAROP	31	5	Δ
NOL	wagon	BALAP	18	6	χ
RUH	office	KUPOD	76	7	∘
BEK	kitchen	MEDON	42	8	-
				9	γ
Example pairs					
JAX	uncle	NARES	71	A	/
ZAD	leader	CAROM	47	B	*
WIQ	zero	FIMUR	90	C	λ

[a] Pairs selected from list used by Martin and Saltz (1963).

[b] Paralogs selected from Noble (1953); numbers from random number table.

make booklets of 100 and 150 pages, respectively.

At the bottom of each page was a line on which the subject was instructed to mark a plus (+) if he thought the stimulus was "correct" or a zero (0) if he thought it was "wrong." He then turned the page and checked his answer against the correct answer printed on the back of the page. On the last page of the booklet he was asked to write the basis for his responses, to verbalize the concept, if he could.

The instructions to the subjects stated that they would be guessing at first, but that as they went along, they should be able to get more and more correct answers, and that they were to try to get as many correct as possible. A subject's learning performance was scored in terms of the number of correct responses for successive blocks of 10 stimuli.

Table 2 presents the pairs used in the paired-associate learning tasks. All three tasks were presented in modified teaching machines using the anticipation method. Thus the subject turned the knob of the machine until the stimulus member of a pair appeared in the opening of the machine. If he knew the response member of the pair, he wrote it on the paper, then advanced the paper so that the stimulus and the subject's response moved under a plastic window and the correct response appeared in the opening. The experimenter set the pace, telling the subjects when to turn to the stimulus, etc. Approximately three seconds were allowed for the stimulus members of the numerical and figural tasks and four seconds for the verbal task in order to provide time for the subjects to write their responses. Approximately two seconds were allowed for the response members of all pairs. Each set of pairs was presented 15 times in different, randomized order, with approximately 10 seconds between repetitions. The subject's score was the number of correct anticipations on each repetition after the first presentation.

The lists used in the verbal and numerical rote-memory tasks are given in Table 3. The lists were each presented 12 times in the same order, at a rate of approximately one word or number every two seconds, with approximately 10 seconds between presentations. Each presentation was followed by a free-recall test. A subject's score was the number of words or numbers recalled correctly on each test, without regard to order of recall.

In the figural rote-memory task, 16 nonsense figures were presented for study for 15 seconds. Immediately following the study was a recognition test in which 16 groups of five nonsense figures were presented, each group containing one of the original figures. The study and test were repeated 12 times, and a subject's score was the number of figures correctly identified on each test.

Ability Tests

The abilities selected for measurement were chosen on the basis of a priori judgment of relevance and the availability of suitable tests. The tests, taken from the Kit of Reference Tests for Cognitive Factors by French, Ekstrom, and Price (1963), were:

Ma: Associative Memory (3 subtests)
Ms: Memory Span (2 subtests)
N: Number Facility (3 subtests)
P: Perceptual Speed (3 subtests)
R: General Reasoning (2 subtests)
V: Verbal Comprehension (2 subtests)

Also included as measures of ability were a Kuhlman-Anderson intelligence score and scores on six subtests of the Stanford Achievement Battery (SAB). The six subtests were:

SAB–PM: Paragraph Meaning
SAB–WM: Word Meaning
SAB–SP: Spelling
SAB–LA: Language
SAB–AR: Arithmetic Reasoning
SAB–AC: Arithmetic Computation

Subjects

The entire sixth grade (135 students in five different classrooms) of an urban New Jersey school participated in the study. Complete information was obtained for 102 children, and all analyses are based upon only those 102.

Procedure

The learning tasks and ability tests were administered in eight sessions, each approximately one hour long. The sessions occurred in the mornings and afternoons of four consecutive days. They were administered in the regular classrooms by psychology graduate students. The distribution of tasks and tests within sessions was determined randomly with the restriction that no learning tasks of the same type or material and no tests of the same ability were included in the same session. There was one exception to this rule; the two memory-span subtests were prerecorded and, to avoid the use of five tape recorders on two days, were administered as the first and last tests of the same session.

In scoring the learning tasks, liberal standards were used with regard to spelling. In addition, in all cases where appropriate, the students were told that there was no penalty for guessing, that if they thought they knew an answer but were not sure to go ahead and mark it.

TABLE 3

LISTS OF WORDS AND NUMBERS USED IN VERBAL AND NUMERICAL ROTE MEMORY LEARNING TASKS

Rote memory—verbal[a]	Rote memory—numerical[b]
MODERN	83
PRETTY	60
GOLDEN	94
FAMOUS	97
GENTLE	13
BETTER	48
COMMON	92
SIMPLE	78
HAPPY	56
NARROW	52
HEAVY	43
LITTLE	36

[a] AA, 2-syllable adjectives from Thorndike & Lorge (1944).

[b] From random number table.

Analysis

Each learning task was scored in terms of the number of correct responses in successive blocks or trials. The Subjects × Trials matrix for each task was analyzed separately according to a method developed by Tucker (1960). The characteristic roots and vectors of the crossproduct matrix formed from the raw score matrix were found and a decision made as to how many were significant.

The factor scores or coefficients of the subjects on these unrotated factors were then calculated. The scores obtained in this way indicate the contribution of each component to each person's performance in that particular learning situation. In other words, the first step determines how many factors are involved in learning a given task; the second step determines the importance or weight of each for a given individual. The scores for the individual were assumed to describe his learning performance.

The individual's scores on all of the learning tasks were then combined with his achievement and ability scores, and the resulting correlation matrix was factor analyzed. Seven factors were extracted, and rotated to an equamax solution.

RESULTS

Learning Tasks

Figure 2 presents the learning curves averaged across the 102 subjects for each task. Learning was not as complete as desirable for the concept-formation tasks, and the number of correct verbalizations of the concepts was correspondingly small, being 20, 41, and 26 for the verbal, numerical, and figural tasks respectively. Learning in the other tasks appears satisfactory.

An indication of the reliability of the factor scores on the learning tasks is provided by the final communality estimates determined by an iterative, refactoring procedure. These estimates provide a lower bound for the reliability and are presented in Table 4. Those for the first component of each task are reasonably high, comparing favorably with those of the ability tests. Those of components after the first are lower, indicating either lower reliability or greater specificity of the component.

Ability Tests

The ability tests were difficult for the subjects, the sixth grade being the lower limit of the range recommended for the tests. Corrected split-half reliabilities and final communality estimates for the ability measures are presented in Table 5.

Learning and Reference Factors

The loadings of the learning-task components and the ability measures on each of the seven factors rotated to an equamax solution are presented in Table 6. The factors were interpreted as follows:

Factor I: Verbal ability. The first factor is interpreted as a verbal-ability factor, and it is common to both the learning tasks and the ability measures.

Factor II: Reasoning ability. All of the reasoning tests have their highest loadings on this factor, and the figural paired - associates and rote - memory tasks have high loadings on it. The verbal and numerical tasks do not.

Factor III: Speed. The common element of the variables loading on this factor seems to be that they involve simple operations which almost everyone can do correctly, given enough time. Accordingly, it is called a speed factor. It is apparently not useful in describing performance in the learning tasks.

Factor IV: Rote-memory ability. This factor is involved in the three associative-memory subtests and in all of the paired-associate and rote-memory learning tasks except figural rote memory. It is interpreted to be a rote-memory factor common to both the ability measures and the learning tasks.

Factor V: Nonverbal learning. This factor is primarily defined by rote-memory tasks involving either numbers or figures. It is interpreted as a nonverbal rote-learning factor distinct from the ability measures.

Factor VI: Concept formation. All of the concept-formation tasks load on this factor and on no other, and no other variable has an appreciable loading on it. Concept formation as defined by the tasks used is not related to the other abilities measured or to the other learning tasks.

Factor VII: Verbal learning. This factor is defined primarily by learning tasks involving verbal material. It is interpreted as a verbal learning factor, independent of the abilities measured.

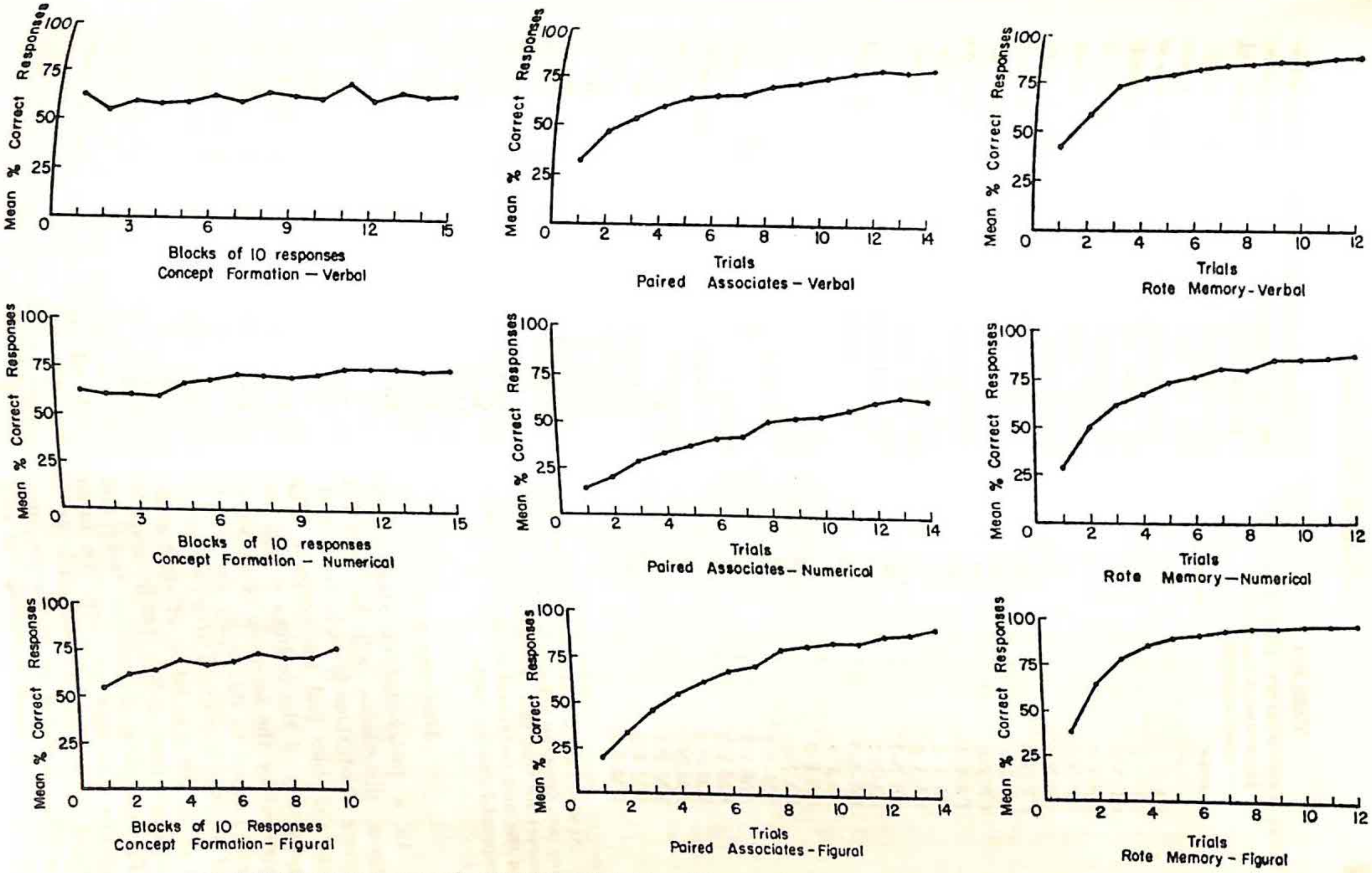


FIG. 2. Trial means expressed as percent of total possible correct responses for each of the nine learning tasks.

TABLE 4

FINAL COMMUNALITY ESTIMATES FOR LEARNING VARIABLES

Variable	Communality
CF-V-1	.34
CF-V-2	.11
CF-V-3	.19
CF-V-4	.04
CF-N-1	.48
CF-N-2	.12
CF-N-3	.16
CF-F-1	.43
CF-F-2	.04
PA-V-1	.55
PA-V-2	.22
PA-V-3	.25
PA-N-1	.60
PA-N-2	.33
PA-N-3	.07
PA-F-1	.46
PA-F-2	.43
PA-F-3	.17
RM-V-1	.57
RM-V-2	.30
RM-V-3	.18
RM-N-1	.51
RM-N-2	.28
RM-F-1	.60
RM-F-2	.37
RM-F-3	.37

Note.—The numbers following the learning task designation indicate the component curves for that task. For example, CF-V-1 is the first component curve of the verbal concept-formation task.

DISCUSSION

In a factor-analytic study, only those dimensions will be revealed along which there is variation in both the variables and the subjects. The generality of the findings is thus restricted by the sampling of measures and people. In this study the sample of subjects is assumed to be representative of the population of sixth-grade children from urban public schools. The sampling of learning tasks and ability measures was, on the other hand, intentionally restricted. It is possible, however, to draw some interesting conclusions regarding the relationships between learning and abilities and learning in different situations.

In the first place, learning is related to measured abilities. With the exception of the concept-formation tasks every learning task was related to one or more of the abilities measured. Furthermore, the abilities were in general appropriate to the a priori classification of the tasks. More specifically, tasks involving words had loadings on verbal factors; tasks involving numbers and figures loaded on nonverbal factors; and paired-associate and rote-memory tasks had loadings on memory factors.

The conclusion that learning and

TABLE 5

RELIABILITIES AND COMMUNALITIES FOR REFERENCE VARIABLES

Variable	Reliability	Communality
IQ	.87[a]	.70
SAB-PM	.90[a]	.82
SAB-WM	.92[a]	.78
SAB-SP	.93[a]	.67
SAB-LA	.81[a]	.58
SAB-AR	.89[a]	.68
SAB-AC	.87[a]	.59
Ma-1	.76	.42
Ma-2	.79	.46
Ma-3	.81	.73
Ms-1	.58	.31
Ms-3	.68	.57
N-1	.91	.73
N-2	.92	.64
N-3	.90	.73
P-1	.81	.29
P-2	.79	.42
P-3	.88	.49
R-1	.44	.50
R-4	.44	.43
V-1	.62	.58
V-2	.68	.42

[a] Reliability coefficients marked taken from the test manual. Coefficients not marked are split-half coefficients corrected for full length with the Spearman-Brown formula.

TABLE 6

ROTATED FACTOR MATRIX

Variable	Factor						
	I	II	III	IV	V	VI	VII
IQ	47	52[a]	40	06	15	15	03
SAB-PM	69[a]	48	28	06	−06	08	13
SAB-WM	77[a]	37	17	03	−07	00	13
SAB-SP	55[a]	24	39	25	20	−04	25
SAB-LA	69[a]	22	15	13	07	08	10
SAB-AR	42	59[a]	34	12	07	04	12
SAB-AC	28	51[a]	40	−11	16	22	09
Ma-1	09	10	18	54[a]	22	08	17
Ma-2	19	17	04	60[a]	14	04	13
Ma-3	33	14	27	51[a]	27	−08	44
Ms-1	44[a]	06	−03	−12	24	18	03
Ms-3	63[a]	−25	05	09	22	17	−17
N-1	26	23	77[a]	−01	11	04	06
N-2	15	32	59[a]	22	32	06	−03
N-3	19	03	80[a]	15	15	06	−06
P-1	01	19	30[a]	16	23	28	−09
P-2	10	06	56[a]	16	18	16	11
P-3	10	46[a]	41	20	04	14	−18
R-1	34	55[a]	12	−05	08	−01	23
R-4	29	48[a]	24	17	04	−03	17
V-1	63[a]	26	17	03	−24	02	18
V-2	55[a]	21	23	06	−11	07	−06
CF-V-1	13	05	−02	−05	15	52[a]	14
CF-V-2	09	11	21[a]	−08	07	12	15
CF-V-3	17	−01	−21	16	−08	25[a]	−14
CF-V-4	04	−02	−03	16[a]	06	03	−05
CF-N-1	−07	14	−04	08	−10	66[a]	09
CF-N-2	−07	−10	−18	−04	06	−25[a]	−06
CF-N-3	−14	08	−03	−03	−25[a]	16	19
CF-F-1	−06	−23	03	06	07	58[a]	17
CF-F-2	−01	04	18[a]	01	00	−09	04
PA-V-1	40[a]	25	22	30	12	24	34
PA-V-2	−26	−02	03	−01	−09	−23	−31[a]
PA-V-3	−25	−17	−13	−26[a]	−21	05	18
PA-N-1	−03	24	01	39	37	16	48[a]
PA-N-2	09	10	−04	−53[a]	04	−11	14
PA-N-3	02	07	05	−03	02	−05	−25[a]
PA-F-1	−04	33	15	43[a]	27	22	17
PA-F-2	−15	27	−06	51[a]	20	13	−08
PA-F-3	03	14	−18	−01	−22[a]	−16	−21
RM-V-1	39	13	34	18	01	07	49[a]
RM-V-2	04	10	16	49[a]	−12	−07	10
RM-V-3	04	−02	−01	18	09	−09	−36[a]
RM-N-1	−02	16	21	33	57[a]	03	08
RM-N-2	−06	−04	08	25	42[a]	03	−15
RM-F-1	03	64[a]	03	01	44	05	−03
RM-F-2	00	55[a]	03	22	−01	09	−07
RM-F-3	−06	08	02	−15	57[a]	01	10

Note.—Decimals are omitted.

[a] The highest loading for a variable.

measures of abilities are related is not new, having been reached both in the field of motor tasks (Adams, 1957; Fleishman, 1960; Fleishman & Hempel, 1954, 1955) and in the field of cognitive tasks (Allison, 1960; Games, 1962; Stake, 1961). The present results are corroborative of these previous findings and extend their generality to learning tasks more representative of those found in other psychological literature.

The second conclusion is that there are learning factors which are independent of the abilities measured. Three such factors were found: The concept-formation factor and the verbal and nonverbal learning factors.

Both Stake (1961) and Allison (1960) also found learning factors, but in neither case did the factors distinguish between verbal and nonverbal tasks. Games (1962) also found two or possibly three remaining factors after rote- and span-memory factors had been partialed out of the learning performances, but he did not interpret them as learning factors.

That the distinction between verbal and nonverbal learning factors is neither obvious nor trivial is indicated both by its failure to appear in the studies cited and by the distinctions which were possible but which did not occur in the present study. For example, the paired-associate and the rote-memory tasks did not define separate factors; nor did the numerical and figural materials separate.

It is possible that the learning factors represent abilities which were not included in these measures, but because of the regularity with which factors additional to ability factors seem to be found, it seems unlikely that adding new ability measures will eliminate them. Rather it seems that there are processes involved in active, repetitive learning situations which are independent of the abilities which can be measured in a static testing situation. Interest and motivation might be contributing factors. Gaudry and Champion (1962) have suggested that anxiety (as measured by the Taylor Manifest Anxiety Scale) affects the limit of learning.

The third major conclusion is that learning in one situation is related to learning in others. In other words, while there does not appear to be a general learning factor, it does appear that there are fewer learning factors than learning tasks. This intertask relationship exists both in the ability and learning factors. Each task loads on a unique combination of factors, but the factors are common to other tasks.

The fourth and final conclusion is that the concept-formation tasks used are unrelated to the other learning tasks and to the abilities measured. It is not clear why this should be so. One possibility is that the tasks called concept formation are rather specialized tasks requiring fairly specific abilities which may or may not be the same as those required in the usual scholastic or intellectual tasks thought of as "concept formation." That is, whether the lack of interrelationships lies in the battery of ability tests or in the specially devised learning tasks or both remains at present a puzzle.

REFERENCES

ADAMS, J. A. The relationship between certain measures of ability and the acquisition of a psychomotor criterion response. *Journal of General Psychology*, 1957, 56, 121–134.

ALLISON, R. B. Learning parameters and human abilities. Office of Naval Research Technical Report. Princeton, N. J.: Educational Testing Service, 1960.

BERG, E. A. A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology*, 1948, 39, 15–22.

FLEISHMAN, E. A. Abilities at different stages of practice in rotary pursuit performance. *Journal of Experimental Psychology*, 1960, **60**, 162–171.

FLEISHMAN, E. A., & HEMPEL, W. E., JR. Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika*, 1954, **19**, 239–252.

FLEISHMAN, E. A., & HEMPEL, W. E., JR. The relation between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology*, 1955, **49**, 301–312.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Kit of reference tests for cognitive factors*. (Rev. ed.) Princeton, N. J.: Educational Testing Service, 1963.

GAMES, P. A. A factorial analysis of verbal learning tasks. *Journal of Experimental Psychology*, 1962, **63**, 1–11.

GAUDRY, E., & CHAMPION, R. A. The identification of learning-curve parameters with intelligence and motivation. *Australian Journal of Psychology*, 1962, **14**, 87–93.

GUILFORD, J. P. Human abilities. *Psychological Review*, 1940, **47**, 367–394.

HUSBAND, R. W. Intercorrelations among learning abilities: I. *Journal of Genetic Psychology*, 1939, **55**, 353–364.

HUSBAND, R. W. Intercorrelations among learning abilities: IV. Effect of age and spread of intelligence upon relationships. *Journal of Genetic Psychology*, 1941, **58**, 431–434.

MARTIN, C. J., & SALTZ, E. Serial versus random presentation of paired-associates. *Journal of Experimental Psychology*, 1963, **65**, 609–615.

McGEOCH, J. A., & IRION, A. L. *The psychology of human learning*. (2nd ed., revised by A. L. Irion) New York: Longmans, Green, 1952.

NOBLE, C. E. The meaning-familiarity relationship. *Psychological Review*, 1953, **60**, 89–98.

STAKE, R. E. Learning parameters, aptitudes, and achievements. *Psychometric Monographs*, 1961, No. 9.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University, 1944.

TUCKER, L. R. Determination of generalized learning curves by factor analysis. Office of Naval Research Technical Report. Princeton, N. J.: Educational Testing Service, 1960.

TYLER, L. E. *The psychology of human differences*. (2nd ed.) New York: Appleton-Century-Crofts, 1956.

WOODROW, H. The ability to learn. *Psychological Review*, 1946, **53**, 147–158.

(Received November 15, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 4, 230–239

ABILITY TO MATCH INFORMATION WITHIN AND BETWEEN AUDITORY AND VISUAL SENSE MODALITIES AND SUBSEQUENT READING ACHIEVEMENT[1]

SIEGMAR MUEHL
University of Iowa

AND

SHIRLEY KREMENAK
Mental Health Institute
Independence, Iowa

To determine the relation between prereaders' abilities to integrate sensory information and reading achievement, 4 matching tasks, involving visual (V) and auditory (A) dot-dash patterns were given to 119 1st-grade Ss in September. The tasks included V-V, V-A, A-V, and A-A matching. School-administered tests included IQ, readiness, and end-of-year reading. Extreme reading groups were selected and matched in IQ. Multiple-regression analyses using extreme group-reading performance as the criterion showed that: (a) When the 4 matching tasks were used as predictors, A-V and V-A tasks made significant contributions to predicting reading, and (b) When the 4 matching tasks were combined with reading-readiness subtests only letter naming made a significant contribution. The relation between A-V matching and letter naming was examined and implications discussed.

Birch and Belmont (1964) have stressed the importance of auditory-to-visual sense integration in learning to read. Using this conceptual framework, one can analyze the beginning reading process with respect to the child's need to make four different kinds of auditory and visual sensory integrations. Auditory and visual discriminations, which are basic to the reading process, are integrations *within* a sense modality. Auditory discriminations require the child to order and relate sound patterns organized on a temporal basis. Visual discriminations require the child to order and relate graphic patterns organized on a spatial basis. Two kinds of integrations *between* the auditory and visual senses are required. Preparatory to reading, the child must relate auditory patterns in speech, which are temporally ordered, to the spatially ordered visual patterns in print. To actually read, he must reverse the process by responding to the printed visual patterns with appropriate sound sequences.

By the time the child is six and is introduced to reading, his experience has equipped him with considerable skill in integrating information *within* and *between* the auditory and visual senses. He can speak a language, which assumes an ability to make auditory integrations; he can discriminate objects, which assumes visual integrations; and he can name objects if asked or point to them if named, which assume visual-to-auditory and auditory-to-visual integrations, respectively.

Are these experience-acquired skills related to the more specific demands of learning to read? Are they independent of IQ? Research has shown a positive relation between auditory discrimination ability for sounds in words and reading achievement independent of both IQ and age (Bond, 1935; Goetzinger, Dirks & Baer, 1960; Wepman, 1960). The relation between visual discrimination ability and reading appears to be dependent upon the stimuli used and the responses required in discrimination testing. Gates (1926) reported a substantial positive relation between word matching and middle-grade reading independent of IQ but only a slight association when discrimination ability was tested using non-verbal materials. Goins (1958) and Barrett (1965) showed that both verbal and nonverbal discrimination tasks predicted first-grade reading independent of IQ. However, in these studies, the so-called "visual discrimination" tasks involved responses of copying and letter naming as well as matching. Studies employing discrimination training with nonverbal (Goins, 1958) and verbal-graphic materials (Muehl, 1960; King, 1964) showed reading facilitation in beginners only following training with verbal materials.

---

[1] Report based on data gathered in second author's PhD dissertation, University of Iowa, 1965.

Research concerned with the ability to relate information between the auditory and visual senses and reading has shown consistent evidence of a positive relationship independent of IQ, age, and the methods and stimuli used in testing (Birch and Belmont, 1964, 1965; Katz and Deutsch, 1963; Walter and Doan, 1962; Williams, 1965).

The purpose of the present study was to investigate these four sensory integration skills using the same subjects and a series of tasks employing comparable stimulus materials and procedures. Two questions were asked: (a) What is the ability of a prereader to integrate or match information *within* and *between* the auditory and visual senses? (b) What is the relation between these matching skills and measures of reading readiness and subsequent reading achievement?

METHOD

Subjects

A sample of 119 subjects (*S*s) was selected from six first-grade classrooms in two elementary schools in Iowa City. For the most part *S*s came from middle-class homes. The *S*s repeating a grade or with known visual or auditory anomalies were not included in the sample. CA ranged from 6-0 to 7-0 with a mean CA of 6-4. Following the beginning-of-the-year testing, 11 *S*s transferred out of the school system leaving a total of 108 *S*s for analyses involving the end-of-year tests.

Testing Materials and Procedures

Matching tasks: Four serial- or sequential-type matching tasks were used. When a visual stimulus was matched with a visual stimulus, the task was designated V-V; auditory with auditory, A-A; auditory with visual, A-V; and visual with auditory, V-A. The stimuli consisted of patterns of dots and dashes, a modification of the dot patterns used by Birch and Belmont (1964). The visual stimuli were drawn with India ink on white 3 × 5 cards. The auditory stimuli were presented with an electric telegraph key which was placed behind a screen so that the experimenter's (*E*'s) operation of the key could not be observed by *S*s. Table 1 shows the warm-up and test items in the four matching tasks. Within each of the four tasks shown in Table 1, six stimuli were arrayed in three levels of difficulty as determined by the length of the stimulus patterns. Items 1 and 2 are two-component stimulus pairs; items 3 and 4, three-component; and items 5 and 6, four-component pairs. The warm-up tasks, consisting of stimulus pairs of one and two components, preceded each of the four tasks.

The four matching tasks were arranged in two orders: *S*s in Order I received V-V, V-A, A-V and A-A; *S*s in Order II received V-V, A-V, V-A, and A-A. This attempt at partial counterbalancing was indicated by a pilot study. The pilot data showed that performance on task pairs containing an initial visual stimulus was about the same whether the tasks were presented early or late in the sequence. However, performance on task pairs with an initial auditory stimulus, particularly the A-A task, tended to improve when given later in the sequence. Since optimal performance was desired, the A-A task, which the pilot data showed most difficult, regardless of position, was placed last. The A-V and V-A were counterbalanced to secure a check on fatigue and warm-up factors that might be operating in the main investigation. *S*s were assigned in rotating order to the two groups.

Incentives were colored stickers and name cards. The *S* picked a colored card

TABLE 1

WARM-UP AND TEST ITEMS FOR FOUR MATCHING TASKS

			Response[a]				Response
Visual-Visual Warm-Up				Auditory-Visual Warm-Up			
1.	–	.	d	1.	.	.	s
2.	.	.	s	2.	– .	– –	d
3.	– .	– –	d	3.	. .	. .	s
4.	. –	. –	s	4.	– .	. –	d
Test				Test			
1.	. .	. –	d	1.	– –	– .	d
2.	. –	– .	d	2.	. –	. –	s
3.	. . –	. – .	d	3.	. . –	– – .	d
4.	. – .	. – .	s	4.	– – .	– . .	d
5.	. . – .	. . – .	s	5.	. . – –	. – – –	d
6.	. – – .	. – – .	s	6.	. – . –	. – . –	s
Visual-Auditory Warm-Up				Auditory-Auditory Warm-Up			
1.	.	–	d	1.	.	–	d
2.	–	–	s	2.	–	–	s
3.	. .	. .	s	3.	. –	. .	d
4.	. –	– .	d	4.	– .	– .	s
Test				Test			
1.	– .	– .	s	1.	. –	. –	s
2.	. .	. –	d	2.	– .	– –	d
3.	. – –	. – –	s	3.	. . –	. . –	s
4.	. – .	. . –	d	4.	– . –	– . –	s
5.	– . – .	– . . –	d	5.	. . . –	. . – –	d
6.	– . . –	– . . –	s	6.	– . – –	– – . –	d

[a] d = different; s = same.

before the task began. The *E* printed *S*s' name on the card. Midway through the task and at the end *S* was allowed to select some stickers to place on his card to take back to class with him. All *S*s got the same number of stickers.

The matching tasks were individually administered by the second author in a small room provided in each school. After the child had been told about the incentive materials, *E* introduced *S* to the stimulus materials and the nature of the task. An introduction preceded the warm-up trials for each of the tasks. During the introductory period, *S* was given the label of "dot" for (.) and "dash" or "line" for (–). He was taught to order the stimuli from left to right by being shown that each pattern must begin and end the same way in order to be considered the "same."

The function of each set of warm-up trials was to give *S* practice in making same-different judgments with similar stimuli and procedures used in the following matching task. Thus, in the A-V task, *S* would first *hear* a pattern of dots and dashes, then *see* on a card a pattern of dots and dashes. He would then be required to make a same-different comparison between the patterns. When the first pattern was visual, the card was always removed before the second pattern in the pair was presented. Thus the matching procedure was sequential in all tasks.

The *E* took *S* through the four warm-up trials for a given matching task once, showing *S* why he was wrong when he made an incorrect judgment. Immediately following the warm-up trials, *E* gave the six test trials.

At the end of each of the four tasks, *S* was told he was doing fine. The *S*'s score was the number of correct same-different responses.

Testing started September 14, 1964, before pupils began formal reading, and was completed on September 28, just as the top group in all classes started reading instruction. Testing time ran about 25 minutes per child.

Reading Readiness Test: In September, the classroom teachers administered *The Harrison-Stroud Reading Readiness Profiles* (Harrison & Stroud, 1956) as part of a school-wide testing program. The test includes seven subtests: Using symbols, Making visual discriminations (attention controlled and uncontrolled), Using the context, Making auditory discriminations, Using context and auditory clues, and Giving names of letters.

Intelligence Test: *The Lorge-Thorndike Intelligence Test, Primary Battery* (Lorge & Thorndike, 1957) was administered to the first-grade classes in March, 1965. Four of the classes were given the test by the teachers, the remaining two classes by the second author.

Reading Tests: In May, the classroom teachers administered *The Metropolitan Achievement Tests, Primary Battery* (Durost, 1959) as part of the school-wide testing program. The reading subtest served as the criterion for reading achievement.

RESULTS

Matching Tasks: The matching tasks were presented in two partially counterbalanced orders. The performance results indicated that the A-V task means were similar regardless of order. The same was true for the V-A means. Therefore, the two orders were collapsed for the subsequent analyses. The matching data were analyzed by means of a four-factor analysis of variance (Lindquist, 1956). The four factors included two within and two between factors. The within factors were the four matching tasks and the three difficulty levels as determined by stimulus length. The between factors consisted of two age levels and sex. The age levels consisted of a low (CA 6-0 to 6-5) and high (CA 6-6 to 7-0) group.

To secure proportionality for the analysis, 13 *S*s, 6 males and 7 females, chosen at random, were dropped. The remaining 106 *S*s were grouped into sex-age categories for purposes of the analyses: young boys ($N = 32$); older boys ($N = 32$); young girls ($N = 21$); and older girls ($N = 21$).

TABLE 2

SCORE DISTRIBUTIONS, MEANS AND STANDARD DEVIATIONS FOR FOUR MATCHING TASKS

Number correct	Task			
	V-V	V-A	A-V	A-A
1	0	0	0	1
2	0	1	4	4
3	0	5	11	31
4	6	12	24	35
5	23	42	29	23
6	77	46	38	12
Total *N*	106	106	106	106
M	5.67	5.20	4.81	4.05
SD	.58	.87	1.15	1.08

Note.—Collapsed over difficulty level, age, and sex.

Table 2 shows the distribution of correct responses, means, and *SD*s in the four matching tasks for these 106 *S*s. Since the distribution of scores on three of the tasks were markedly skewed the level of significance for the analysis was set at the .025 level. The results of the analysis revealed significant effects associated with matching tasks, stimulus length, and with the matching task-stimulus length interaction ($p < .001$). The matching task means, collapsed over difficulty levels, are presented in Table 2. V-V proved to be easiest, A-A the most difficult, and V-A and A-V intermediate in difficulty in that order. The interaction between matching tasks and stimulus length or difficulty level revealed that an increase in stimulus length had no effect on performance in the V-V task but made the V-A, A-V and A-A tasks more difficult. There were no reliable differences associated with sex or age.

Matching Task versus Reading Performance: To maximize differences in reading ability the reading scores

TABLE 3

INTERCORRELATIONS AMONG FOUR MATCHING TASKS AND HIGH AND LOW READING ACHIEVEMENT ($N = 49$)

Variables	Intercorrelations			
	2	3	4	5[a]
V-V task	.16	−.04	.19	.16
V-A task		.27	.40*	.39*
A-V task			.25	.52*
A-A task				.28
Reading				

[a] r_{bis}.

* $p < .05$.

were dichotomized into High and Low groups. The standard scores on the Metropolitan reading test have a mean of 50 and *SD* of 10. The standard scores for the 108 *S*s available for end-of-year testing ranged from 30 to 67 with a mean of 56.83. The Low group consisted of *S*s whose standard scores were 52 or lower, or the bottom 23% of the *S*s. Since a standard score of 52 corresponds to the 60th percentile on the Metropolitan norms, a few *S*s in the Low group were actually above average in ability by this standard. The majority, however, ranged from the 25th to the 50th percentile. The High group included *S*s with a standard score of 65 or higher, or the top 22% of *S*s. The standard score of 65 corresponds to the 95th percentile on the Metropolitan norms. Since it was desirable to control IQ differences between the High and Low groups, the groups were matched using the Lorge-Thorndike scores.[2] In the process, four Low IQ *S*s were dropped. The final *S* totals for the High and Low groups were 24 and 25, the mean IQs, 119.7 and 117.6, and the mean CAs, 6-4 and 6-5, respectively. The group differences in IQ and CA were not significant. The High group contained eight boys and 16 girls; the Low, 17 boys and eight girls. This distribution by sex differed reliably from chance ($\chi^2 = 4.58$, $p < .05$).

To determine the relation between prereading matching task performance and extreme group reading performance at the end of the year, a discriminant function analysis was used. This analysis resembles the multiple regression analysis except that it permits a dichotomized dependent variable. The analysis program[3] deletes those independent variables contributing least to the multiple correlation with no loss of predictive power. In the first analysis, the dichotomized reading group scores (1 for High; 0 for Low) served as the criterion variable. The matching task scores served as predictor variables.

Table 3 shows the intercorrelations among the matching tasks and reading achievement. Table 4 shows the summary of the discriminant function analysis. The four predictor variables yielded a final multiple correlation coefficient (R) of .58 with reading. Only A-V and V-A matching contributed significant *beta* weights to the final R.[4] In Table 5, the High and Low readers are divided into high and low performance on the A-V matching task. The table shows that only one of 23 High group readers had a low A-V score, whereas over half of the

---

[2] The Lorge-Thorndike score was not included as one of the predictor variables in the subsequent analyses since the testing date was in March as compared to September for the matching and reading readiness tasks. Results including the IQ as a predictor variable included in subsequent footnotes.

[3] Program developed by Bechtoldt at University of Iowa Computer Center for the IBM 7044, using Fortran IV.

[4] An analysis using all 108 *S*s which included IQ and the four matching tasks as predictors yielded a final R of .45. Significant *betas* were reported for V-A (.23) and A-V (.34) matching.

TABLE 4

DISCRIMINANT FUNCTION ANALYSIS INVOLVING FOUR MATCHING TASKS AS PREDICTORS AND HIGH AND LOW READING ACHIEVEMENT AS CRITERION ($N = 49$)

Predictor variables	Initial betas[a]	Final betas[b]
V-V task	.13	
V-A task	.23	.27*
A-V task	.45*	.44*
A-A task	.05	

[a] $R = .60$.
[b] $R = .58$.
* $p < .05$.

Low readers were low scorers on the A-V task.

To determine the degree of relationship between the matching tasks and the reading-readiness tests, and the contribution of each to predicting reading, a second discriminant function analysis was run. Scores from the seven readiness subtests, plus the four matching tasks served as predictor variables. The dichotomized reading scores again served as the criterion variable. Table 6 summarizes the results of this analysis. With all but two tests deleted (Using context and Giving names of letters), the final R was .83. Of the two, only letter naming had a significant *beta*.[5] Since the letter-naming test made such a large contribution to predicting reading performance, it was of interest to examine the correlations between this variable and the four matching tasks. The V-A and A-V tasks correlated significantly with letter naming in the extreme groups,

TABLE 5

DISTRIBUTION OF HIGH AND LOW READERS ON A-V MATCHING TASK

	Low A-V scores (1-4 correct)	High A-V scores (5-6 correct)	Total
High readers	1	23	24
Low readers	13	12	25
Total	14	35	

TABLE 6

DISCRIMINANT FUNCTION ANALYSIS INVOLVING FOUR MATCHING TASKS PLUS READING READINESS SUBTESTS AS PREDICTORS AND HIGH AND LOW READING ACHIEVEMENT AS CRITERION ($N = 49$)

Predictor variables	Initial betas[a]	Final betas[b]
Harrison-Stroud subtests		
Using symbols	−.02	
Making visual discriminations		
Attention span controlled	.10	
Attention span uncontrolled	−.05	
Using the context	.16	.15
Making auditory discriminations	−.24	
Using context & auditory clues	.24	
Giving the names of letters	.74*	.79*
Matching tasks		
V-V task	.14	
V-A task	−.04	
A-V task	.14	
A-A task	−.01	

[a] $R = .86$.
[b] $R = .83$.
* $p < .05$.

TABLE 7

DISTRIBUTION OF HIGH AND LOW READERS ON LETTER-NAMING SUBTEST

	Giving names of letters		
	Lowest 50% (score ≤ 17)	Highest 50% (score ≥ 18)	Total
High readers	0	24	24
Low readers	13	12	25
Total	13	36	

---

[5] An analysis using all 108 Ss which included IQ, readiness subtests, and the matching tasks as predictors yielded simple correlations with reading as follows: Readiness subtests—Using symbols (.33), Visual discrimination controlled (.28), Visual discrimination uncontrolled (.24), Using context (.19), Auditory discriminations (.32), Using context and auditory clues (.27), Letter naming (.57); Matching tasks—V-V (.10), V-A (.30), A-V (.40), A-A (.25); IQ (.10). The final R was .66 with significant *betas* reported for Using symbols (.17), Letter naming (.46), V-A (.16) and A-V (.17).

.40 and .53 respectively. In Table 7, the High and Low readers are divided into groups with high and low performance on the letter-naming subtest. As can be seen, the distribution is almost identical to the one obtained using A-V matching-task performance (see Table 5).

DISCUSSION

The first purpose of the present study was to investigate prereaders' ability to match information *within* and *between* the auditory and visual sense modalities. The results revealed marked differences in first-graders' ability to make "same-different" comparisons between dot-dash patterns presented as auditory, visual, and cross-modal pairs (see Table 2). Matching visual pairs was easy for most children. Out of six test items, the majority of children made a perfect score. None got fewer than four correct. Matching auditory pairs was the most difficult task. Here, about one-third scored at chance or less. Matching visual-to-auditory and auditory-to-visual pairs was intermediate in difficulty, with the latter task proving the more difficult of the two.[6]

The second purpose of the study was to investigate the relationship between these four sensory matching skills and reading readiness and subsequent reading achievement. When extreme reading groups (based on end-of-year tests) were equated in IQ, the analysis showed that beginning-of-the-year ability to match visual-to-auditory (V-A) and auditory-to-visual (A-V) pairs made significant contributions to predicting reading. Table 5 shows that only one of 14 *S*s who had a low A-V score at the beginning of the year appeared in the high reading group at the end of the year. Of those *S*s with high A-V scores, twice as many ended up in the high reading group. Thus, for children used in this study, the general ability to relate information from the auditory to the visual sense, with IQ controlled, was markedly associated with later reading achievement. However, since some of the better A-V performers fell into the low reading groups, the skill is clearly not the only variable involved. The relationship of the ability to relate information between the auditory and visual sense and reading found in this study is consistent with the results of previously cited research.

Ability to match visual pairs did not contribute to predicting reading achievement. This result was to be expected since the task was so easy as to fail to yield a distribution of performance differences. Auditory matching ability yielded the best distribution of performance scores of the four matching tasks (see Table 2). Yet individual differences in this skill made no independent contribution to predicting reading achievement. This finding is inconsistent with the research evidence and at odds with pedagogical opinions regarding the importance of this skill in readiness and early reading instruction. The failure to find a relationship in the present study may be due to the use of non-verbal materials. Previous studies (Goetzinger, Dirks and Baer, 1960; Wepman, 1960) used spoken words as auditory stimuli. Reading may well require a more specialized type of auditory discrimination skill developed only with verbal materials. The lack of relationship could also be attributed to the relatively high level of language ability of the children used in the study (as testified to by the IQ and reading scores). It may be that only a given level of auditory discrimination skill is prerequisite for learning to read; and that this level was already present in most of these children as an aspect of their general language skill.

---

[6] Since the elements *within* an auditory stimulus pattern cannot be presented simultaneously as can the elements within a visual pattern, the auditory pattern is probably more difficult to perceive and recall, thus making comparisons involving these stimuli less accurate.

If this second line of reasoning is correct, the results indicate that children with a good language background do not need extensive instructional exercises to make them conscious of differences between sound patterns. Rather, the emphasis should be directed toward making them conscious of auditory patterns *in relation to* visual-graphic patterns. However, this view would also suggest that if children come to school with marked deficiencies in general language ability, then training in auditory discrimination would be important (Deutsch, 1964). Durrell and Murphy's (1953) findings that auditory training was helpful in improving reading achievement only for children poorest in this skill lends support to this view.

When the four sensory matching tasks were combined with the reading readiness subtests to predict end-of-year reading achievement for the extreme reading groups, the ability to name letters made the greatest contribution by far. This result is consistent with findings reported by Durrell (1958) and Barrett (1965). What is involved in the ability to name letters that accounts for the relation with reading? The present findings indicated that, in part, letter-naming ability reflected a more general ability to relate information between the auditory and visual sense modalities. This fact is borne out by the significant correlations between the A-V and V-A tasks and the letter-naming test. However, at the same time, there were marked differences among the correlations of these three tasks with reading achievement. Letter naming correlated .82 with extreme group reading achievement with IQ controlled; the A-V and V-A tasks correlated .52 and .39, respectively. Although test reliability attenuated to some degree the latter coefficients, it is likely that other factors contributed to the additional predictive power of the letter-naming task. For example, the ability to name letters may aid in labeling, which would be assumed to facilitate discrimination and recall of printed words. The sound similarity between letter names and the phonemes these letters represent in words may help in word analysis (Muehl, 1962).

The findings of the present study and similar research appear to have practical implications for the classroom. The evidence points to the need for early identification of children deficient in the ability to integrate information conveyed through the auditory and visual senses. Naming the letters of the alphabet provides a means for screening these children. Although it does not necessarily follow that training in letter naming is the best way of giving children practice in developing the auditory-to-visual integrating ability, there is some evidence that such training may be helpful. Durrell (1958) showed that systematic instruction in letter names and sounds in first grade produced better reading achievement when compared to incidental instruction in this skill.

However, the children in the present study *had* received systematic instruction in letter names and sounds during the latter half of the kindergarten year. Thus, deficiencies in letter-name knowledge existed at first-grade level despite this instruction. Perhaps the ability to relate abstract symbols from the auditory to the visual sense modality (e.g., dot-dash patterns, letters and their names) is in part a maturational function. Birch and Lefford's (1963) findings relating to the emergence of intermodal equivalence for the visual, kinesthetic, and haptic senses support this view. They found growth curves similar to those reflecting anatomic development as in height and weight.

This reasoning would suggest that children who are unable to learn the names of the letters probably will not be able to learn the more complex visual-auditory associations involving whole words. Thus a continuing readiness program should be provided these children until they indicate sufficient maturity to master the letter-naming task.

REFERENCES

BARRETT, T. C. Visual discrimination tasks as predictors of first grade reading achievement. *Reading Teacher*, 1965, **18**, 276–282.

BIRCH, H. G., & BELMONT, L. Auditory-visual integration in normal and retarded readers. *American Journal of Orthopsychiatry*, 1964, 34, 852–861.

BIRCH, H. G., & BELMONT, L. Auditory-visual integration, intelligence and reading ability in school children. *Perceptual and Motor Skills*, 1965, **20**, 295–305.

BIRCH, H. G., & LEFFORD, A. Intersensory development in children. *Monographs of the Society of Research in Child Development*, 1963, **28**, (5, Whole No. 89).

BOND, G. L. The auditory and speech characteristics of poor readers. *Teachers College Contributions to Education*, 1935, No. 657.

DEUTSCH, C. Auditory discrimination and learning: Social factors. *Merrill-Palmer Quarterly*, 1964, **10**, 277–296.

DUROST, W. N. (Ed.) *Metropolitan achievement tests, primary battery*. Yonkers-on Hudson: World Book, 1959.

DURRELL, D. D. First grade reading success: A summary. *Journal of Education*, 1958, **140**, 2–6.

DURRELL, D. D., & MURPHY, H. The auditory discrimination factor in reading readiness and reading disability. *Education*, 1953, **73**, 556–560.

GATES, A. I. A study of the role of visual perception, intelligence and certain associative processes in reading and spelling. *Journal of Educational Psychology*, 1926, **17**, 433–445.

GOETZINGER, C. P., DIRKS, D. D., & BAER, C. J. Auditory discrimination and visual perception in good and poor readers. *Annals of Otology, Rhinology, and Laryngology*, 1960, **69**, 121–136.

GOINS, J. T. Visual perceptual abilities and early reading progress. *Supplementary Educational Monographs*, Chicago: 1958, No. 87.

HARRISON, L., & STROUD, J. B. *The Harrison-Stroud reading readiness profiles*. Boston: Houghton Mifflin, 1956.

KATZ, P., & DEUTSCH, M. Relation of auditory-visual shifting to reading achievement. *Perceptual and Motor Skills*, 1963, **17**, 327–332.

KING, E. M. Effects of different kinds of visual discrimination training on learning to read words. *Journal of Educational Psychology*, 1964, **55**, 325–333.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1956.

LORGE, I., & THORNDIKE, R. L. *The Lorge-Thorndike intelligence tests*. Boston: Houghton Mifflin, 1957.

MUEHL, S. The effects of visual discrimination pretraining on learning to read a vocabulary list. *Journal of Educational Psychology*, 1960, **51**, 217–221.

MUEHL, S. The effects of letter-name knowledge on learning to read a word list in kindergarten children. *Journal of Educational Psychology*, 1962, **53**, 181–186.

WALTER, R. H., & DOAN, H. Perceptual and cognitive functioning in retarded readers. *Journal of Consulting Psychology*, 1962, **26**, 355–361.

WEPMAN, J. M. Auditory discrimination, speech and reading. *Elementary School Journal*, 1960, **60**, 325–333.

WILLIAMS, J. Auditory-visual equivalence of stimuli among normal and retarded readers. Unpublished master's thesis, University of Iowa, 1965.

(Received November 24, 1965)

The American Psychological
Association announces . . .

the second annual publication of the "PROCEEDINGS"

PROCEEDINGS of the 74th Annual Convention of the American Psychological Association 1966*

Contributed Papers of the Divisions of Experimental Psychology, Physiological & Comparative Psychology, Clinical Psychology, Educational Psychology, and Counseling Psychology at the 74th Annual Convention of the American Psychological Association, 1966.

The purpose of *Proceedings* is to facilitate the dissemination of information regarding current work in major fields of psychology by making available detailed abstracts of this research in advance of the convention presentation and journal publication. Because of the value of this undertaking to research, the plans are to make the *Proceedings* a permanent feature of APA Conventions and extend its contents to other APA Divisions. "Proceedings of the 73rd Annual Convention of the American Psychological Association" was issued in 1965 on a trial basis. The favorable reception of that issue indicated that a permanent successor would be worthwhile.

There will be approximately 350 pages of text and 6,000 copies of the publication will be printed and ready for issue in July. This publication will be o archival value and will be available to libraries as well as to individuals.

PRICE: $6.00 per copy ($3.00 to APA members)

AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 Seventeenth Street, N.W., Washington, D. C. 20036

Clip & Mail Today

AMERICAN PSYCHOLOGICAL ASSOCIATION,
1200 17th St., N.W., Washington, D. C. 20036

Please enter my order now for . . . copy(ies) of *Proceedings* (*1966*) to be mailed when printed (approximately 1 July 1966).

NAME . . . .

ADDRESS . . . .

☐ Remittance enclosed ☐ Bill me when issue mailed

---

* Previous announcements have given CURRENT RESEARCH IN PSYCHOLOGY as the title of this publication. However, in the interest of continuity it has been decided to retain the title of "PROCEEDINGS".

RUSSIAN MONOGRAPHS ON BRAIN AND BEHAVIOR

A series of translations of Russian Monographs concentrating upon the historical background of investigations in neurophysiology and conditioned behavior as well as upon certain contemporary advances in these fields.

The translation program was initiated by H. W. Magoun and D. B. Lindsley and has been continued by the latter. The complete series will consist of eleven monographs published in hard-cover book form. To date, three of the eleven monographs have been printed and are available through the American Psychological Association.

Available now . . .

ESSAYS ON THE HISTORY OF PHYSIOLOGY IN RUSSIA

By Kh. S. Koshtoyants.

These essays review broadly the history of the development of physiology as a science in Russia, beginning in the early 18th century and continuing to modern times. They are more than a history for they deal with the personal characteristics and strivings of great scientists who forged teaching and research disciplines in anatomy, physiology, and psychology.

321 pp. $4.00

AUTOBIOGRAPHICAL NOTES

By I. M. Sechenov.

Sechenov, revered as the father of physiology in Russia, was born in 1829. His scientific and academic pursuits, though always stimulating and challenging, were fraught with difficulties due to political interference and harassment. These **Autobiographical Notes,** written about a year before his death in November, 1905, provide a vivid picture of his history.

174 pp. $4.00

ORIENTING REFLEX AND EXPLORATORY BEHAVIOR

Edited by L. G. Voronin, A. N. Leontiev, A. R. Luria, E. N. Sokolov, and D. S. Vinogradova.

A collection of papers read at a Conference organized in February, 1957 by the Institute of Defectology of the Academy of Pedagogical Sciences, RSFSR, and by the Department of Psychology and Higher Nervous Activity of the M.V. Lomonsov Moscow State University. As the varied experiments and points of view in this collection of papers will reveal, a serious attempt has been made to establish the role and significance of the orienting-exploratory reaction, not only in the elementary behavior of lower animal forms, but also in the complex behavior of higher forms, including man and his cognitive functions.

462 pp. $4.00

AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 Seventeenth St., N. W., Washington, D. C., 20036

JOURNAL OF EDUCATIONAL PSYCHOLOGY

Volume 57 October 1966 Number 5

CONTENTS

Validity Judgments of Syllogisms in Relation to Two Sets of Terms........... 239

LAWRENCE T. FRASE

Delay of Gratification as a Function of Characteristics of Social Agents........ 246

MARCIA J. LAWTON

Refinement and Test of the Theory of Fluid and Crystallized General Intelligences........ 253

JOHN L. HORN AND RAYMOND B. CATTELL

Semantic Constraint in Paired-Associate Learning........ 271

WILLIAM D. ROHWER, JR., AND STEVE LYNCH

Size-Discrimination Learning as a Function of Motivation-Hygiene Orientation in Adolescents........ 279

H. CARL HAYWOOD AND THEODORE D. WACHS

Group Structure and Productivity in Problem-Solving........ 287

BRYCE B. HUDGINS AND LOUIS M. SMITH

Predictive Efficiency of Two Multivariate Statistical Techniques in Comparison with Clinical Predictions........ 297

C. E. ALEXAKOS

Comparison of the Effects of Positive and Negative Vicarious Reinforcement in an Operant Learning Task........ 307

RICHARD KELLY

Journal of Educational Psychology
1966, Vol. 57, No. 5, 239–245

VALIDITY JUDGMENTS OF SYLLOGISMS IN RELATION TO TWO SETS OF TERMS[1]

LAWRENCE T. FRASE

University of Illinois[2]

The separate effects of affective and quantifying words upon syllogistic reasoning were studied in a 2 × 3 × 2 × 2 factorial design. Qualifiers rated by 64 Ss on evaluative scales were combined to yield 3 levels of incongruous conclusions; 2 levels of quantifiers were used (universal/particular). 32 undergraduate Ss received relevant syllogistic training; 32 received irrelevant training. Results reflected the operation of 2 independent verbal response systems. Although both affective and quantifying terms produced significant time and error effects, the quantifier "some" produced gross effects and was sensitive to training while responses to affective terms were relatively stable. Moderate levels of incompatibility produced most errors. Initially, Ss judged the validity of highly incompatible conclusions rapidly; after training they spent most time on those conclusions.

Consider the following syllogism:

All good things are fine.
All beautiful things are fine.
Therefore, all beautiful things are good.

Suppose we substitute the word "evil" for the word "good." Will subjects (*S*s) be more likely to conclude that the syllogism is invalid when the conclusion reads, "All beautiful things are evil"? More important, can we combine words (which vary in their evaluative direction and intensity) and systematically manipulate the number of errors or time it takes to make deductive judgments?

Suppose that we change the conclusion "All beautiful things are evil" to "Some beautiful things are evil." Does the conclusion now seem to follow more or less logically from the premises? What is the effect of training upon error and time responses to the formal logical terms "All" and "Some"?

The formal syllogism is a verbal problem-solving task in which two separate classes of terms can be distinguished. The first class of terms (all, some, no, not) determine the structural characteristics of the syllogism which are the sufficient conditions for judging the validity of the conclusion. The second class of terms (syllogistic variables), whether X's, Y's, or meaningful words, have no bearing on the validity of the syllogism, but may influence the correctness of an individual's inference. For instance, the effects of a neutral sentence "All X are Y" can be compared with the effects of emotional statements such as "All communists are torturers" (Lefford, 1946), or statements which produce low or high belief, such as, "All men are women" as opposed to "All men are males" (Janis & Frick, 1943).

Mowrer (1954, 1960) has dealt with the relationship between verbal signs in a sentence as a means for producing transfer of response, and Staats (1959) has presented evidence which demonstrates that nonsense syllables acquire negative or positive evaluative tone when paired with appropriate evaluative response terms (predicates). It is anticipated that when the terms to be associated elicit antagonistic responses, interference results. If such incompatible responses occur in the conclusion of a syllogism, reasoning should be disrupted.

---

[1] This paper is based on the author's doctoral dissertation in education at the University of Illinois and was supported by Office of Naval Research Contract 3985 (04) and funds granted from the University of Illinois Psychology of Classroom Learning Program.

[2] Now at the University of Massachusetts.

In support of the foregoing analysis are data relating to response latency. Berlyne (1960) found that it took significantly longer for *S*s to rate pairs of words on Semantic Differential (SD) scales when the pairs rated were far apart in semantic distance. What Berlyne refers to as "conceptual conflict" is similar to the concept employed in the present experiment, referred to as "incompatibility (*I*)." Solarz (1960) also found supportive data; latencies of arm movements were shorter in response to stimuli which were compatible (elicited positive evaluative ratings) with approach movements. Hence, it is expected that combinations of incompatible words will increase judgment time.

In regard to the formal stimuli, data from previous studies on the syllogism (Sells, 1936; Wilkins, 1928) indicate that *S*s make more errors on *particular* conclusions (some, some . . . are not) than on *universal* conclusions (all, no). Although *S*s have been told the definition of "some" before an experiment (to eliminate any systematic bias that might be introduced), it is not clearly the case that they actually do apply this governing definition. Studies on the effect of training on specific verbal factors in syllogistic reasoning have been absent from the literature. The assumption that a verbal introduction to logical terms overcomes the effects of those terms should be substantiated.

The major interest of this paper is the relative and absolute effect of formal and affective terms upon reasoning both before and after training. In the present study, as opposed to earlier studies, predictions are made on the basis of an associative analysis of combinations of individual words rather than on the basis of entire sentences.

METHOD

Stimulus Words[3]

A total of 44 qualifiers, such as "sinful," "fearful," "successful," which were relatively "active" and "potent," and either "positive" or "negative" in evaluative polarization were selected from norms of English-speaking *S*s (Osgood, Archer & Miron, 1962). These words were to be used as subject terms in the syllogisms. Predicate and middle terms were selected from other SD scales. The aim was to get a large variety of words so that many levels of incongruous word combinations could be formed.

Rating Scales

Four 7-point scales were used to evaluate the stimulus words selected as possible terms for the syllogisms. The scales, good-bad, dirty-clean, nice-awful, and heavenly-hellish, were selected on the basis of reliability from the data of Osgood, Archer, and Miron (1962). Four sets of nonevaluative scales were used as filler items. Four sets of words and scales were presented per page with scale directions alternated to eliminate bias.

Response Incompatibility (I)

The range of possible scores for each word was from 4 to 28. The formula for *I* with affirmative conclusions was the absolute value of the difference between the mean ratings of the words ($\overline{\mathrm{W}}$) comprising the conclusion.

$$I_{\mathrm{A}} = /\overline{\mathrm{W}}_1 - \overline{\mathrm{W}}_2/ \qquad [1]$$

For instance, if "mother" were rated 28 and "evil" rated 4, then the universal affirmative proposition "All mothers are evil" would receive a rating of 24, that is,

$$I_{\mathrm{A}} = /28 - 4/ = 24.$$

---

[3] The complete report of these words and syllogisms used is deposited with University Microfilms, Inc., 313 North First Street, Ann Arbor, Michigan.

In the case where word combinations were to be used in negative propositions the formula was

$$I_N = //\overline{W}_1 - \overline{W}_2/ - 24/. \qquad [2]$$

The proposition "No mothers are evil" would thus receive an I value of zero.

Formulas 1 and 2 are analogous to "pressure of incongruity" used by Osgood and Tannenbaum (1955), except that I is expressed in scale units more convenient for this experiment.

Test Syllogisms

The same syllogisms were used at three levels of I. Each of 8 syllogisms terminating in a *universal* conclusion had its counterpart terminating in a *particular* conclusion. These universal and particular syllogisms were matched in terms of validity and the logical rules they violated. There was a total of 48 test syllogisms, 12 of which were valid. The mean I values for conclusions used at the three levels were 5.8, 13.5, and 20.6. There was no overlap between categories.

As examples, the following two syllogisms have an I value of .71 (low) and differ only in terms of the quantifier used in the conclusion.

All sinful things are cruel
  and all disgusting things are sinful.
Therefore, all disgusting things are cruel.
All sinful things are cruel
  and all disgusting things are sinful.
Therefore, some disgusting things are cruel.

Two moderate I syllogisms ($I = 13.9$) are given below, illustrating a negative conclusion.

Some interesting things are foolish
  and some foolish things are fearful.
Therefore, no fearful things are interesting.
Some interesting things are foolish
  and some foolish things are fearful.
Therefore, some fearful things are not interesting.

Highly incompatible combinations of words ($I = 23.5$) occur in the following syllogism.

No natural things are creative
  and no natural things are successful.
Therefore, no successful things are creative.

Subjects

Sixty-four undergraduate students from the University of Illinois (with no previous training in formal logic) from 28 academic areas were paid $1.00 per hour for participating in the experiment. Grade levels ranged from freshman to senior.

Apparatus

The computer-based teaching machine system SOCRATES (Training Research Laboratory, University of Illinois) was used to present the test syllogisms. See Merrill (1965) for references on this system. The Master I/O (input-output) stations associated with this system provide random access to any one of 1500 35-mm. filmed images. The filmed image is projected on a high-contrast rear-projection screen. A red response button was located on the desk top next to the I/O station which was located in a DuKane language laboratory booth.

Two Master I/O stations were employed for this study. When S reported for a test he was assigned to a machine and the logical meaning of "all" and "some" was explained verbally according to the procedure reported by Sells (1936). The Ss were given a response booklet containing instructions on how to indicate the validity of each syllogism. The instruction booklet also repeated instructions on the logical meaning of "some." When S was ready to begin, or to proceed to a new test item, he pressed the red response button. At this time SOCRATES randomly selected a test item for S, recording the time (on IBM cards) from the illumination of the test frame until the response button was released.

Procedure

The Ss rated the 44 words on SD evaluative scales. The mean evaluative ratings of these words was used to select word combinations at three levels of I. The resulting syllogisms were then filmed.

Five weeks after the initial rating session the same Ss took the syllogistic test over the 48 matched syllogisms. There was no time restriction on their judgments.

One week later the same Ss were randomly assigned to two groups. One group underwent 1½ hours of programmed instruction in syllogistic reasoning in which it was explicitly stated that syllogistic variables (affective terms) have no bearing upon the formal validity of a syllogism. The distinction was drawn between formal validity and the truth of a conclusion. The other group studied an irrelevant task (concerned with the Polish system for symbolizing propositions) for a similar period of time. For this group, no reference was made to syllogistic

TABLE 1

MEANS AND STANDARD DEVIATIONS OF TIME AND ERROR SCORES FOR MAIN INDEPENDENT VARIABLES

Variable	Time in seconds[a]				Errors			
	Pretest		Posttest		Pretest		Posttest	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Training[b]								
Irrelevant	344.65	138.17	255.43	88.83	2.35	1.05	2.27	1.30
Relevant	354.72	121.33	260.41	121.00	2.22	1.19	1.86	1.16
I[c]								
Low	356.51	128.63	253.30	108.79	2.06	1.25	2.02	1.29
Moderate	355.89	145.50	258.17	111.50	2.47	1.33	2.10	1.40
High	336.65	135.42	262.29	115.50	2.33	1.26	2.08	1.34
Conclusions[c]								
Universal	323.32	118.00	243.86	99.32	1.52	1.09	1.54	1.19
Particular	376.05	149.93	271.98	120.00	3.06	1.44	2.59	1.51

Note.—Time and error scores represent the average for 8 syllogisms.

[a] Due to mechanical failure, time records for 4 *S*s from each training group were unavailable.

[b] Time, $N = 28$; Errors, $N = 32$.

[c] Time, $N = 56$; Errors, $N = 64$.

variables or affective terms. Both programmed texts were developed by the same authors (Kearns, Odell, & Zartman, 1964). The *S*s were not informed of the purpose of the present study.

One week later all *S*s retook the syllogistic test of 48 items.

Design

A 2 × 3 × 2 × 2 factorial design (fixed factors) with repeated measures on the last three factors was used; Factor 1: relevant or irrelevant training; Factor 2: low, moderate, or high I conclusions; Factor 3: universal or particular conclusions; Factor 4: pretest versus posttest scores.

RESULTS

Time

There was no significant main effect of I on time. There was a significant interaction between pre-posttest and I ($F = 3.15$, $df = 2/108$, $p < .05$), indicating that *S*s initially made snap judgments on incompatible conclusions, but on the posttest they spent most time on those conclusions. The mean judgment times are reported in Table 1.

Total time was reduced on the posttest ($F = 49.1$, $df = 1/54$, $p < .001$). The means for *particular* and *universal* quantifiers were significantly different ($F = 38.02$, $df = 1/54$, $p < .001$). The formal terms accounted for 27% of the within-subject variance in time scores, while the affective terms accounted for 1% of the variance.[4]

Cumulative plots of time scores indicated no warm-up effects.

Errors

There were significant main effects of both affective and quantifying terms on errors. For I the F was 3.98 ($df = 2/124$, $p < .05$). The F for quantifiers was 141.88 ($df = 1/62$, $p < .001$). Multiple comparisons (Newman-Keuls method) indicated that the moderate level of I differed significantly ($p < .05$) from low I, but the mean at high *I* did not differ significantly from either low or moderate *I* level means (Figure 1). There was a tendency toward reduction in the effect of *I* after training, but the interaction *F* between pre-posttest and *I* was not significant ($F = 1.85$, $df = 2/124$).

---

[4] See Hays (1963) for the estimate of the proportional reduction in variance of Y given X—est. ω^2.

Figure 2 summarizes the effects of *universal* and *particular* quantifiers and the relative influence of training upon errors. Differences between group performance on quantifiers were significant only for *particular* conclusions on the posttest ($p < .01$). The training group-qualifier-pre/posttest interaction *F* was 4.33 ($df = 1/62$, $p < .05$), and the interaction *F* for training group-qualifier was 8.90 ($df = 1/62$, $p < .01$). Although the specific effects of instruction are apparent from Figure 2, over the total test the trained group was not superior to the untrained group. On a later test (not a part of the present analysis) the trained group was superior.

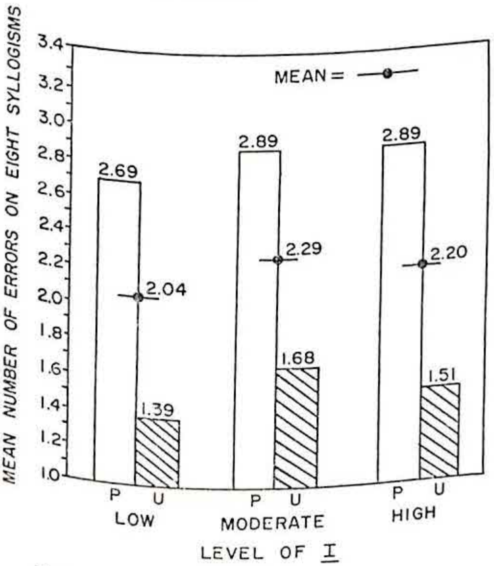


Fig. 1. Mean number of errors at three levels of *I* for syllogisms with universal (U) or particular (P) conclusions. (The bar indicates the average of the universal and particular conclusions combined.)

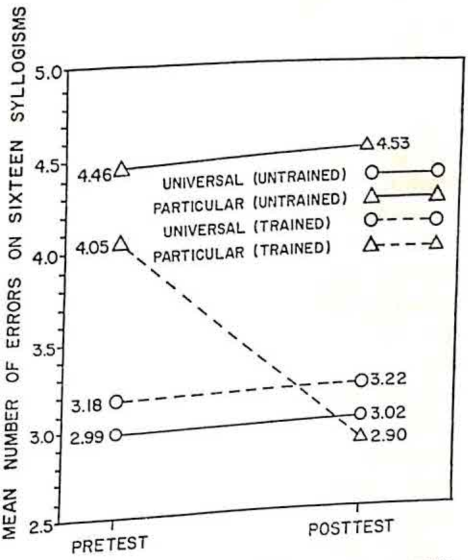


Fig. 2. Effect of training upon errors made on universal and particular quantifiers for the group trained and the group not trained in syllogistic reasoning.

Formal terms accounted for 26% of the within-subject variance in error scores (est. ω^2). Affective terms accounted for 1% of the within-subject variance.

As in the case of time scores, there were no warm-up effects during testing.

Discussion

Results of the present study indicated that presenting instructions on the logical meaning of "some" verbally and in the test instruction booklet alone was not a sufficient condition for that term to effectively control behavior. The group which underwent relevant instruction showed significantly greater improvement on syllogisms with *particular* conclusions. The "atmosphere hypothesis," which was the major thesis of Sells' (1936) early work, consisted of 16 predictions (of what conclusions would be chosen as valid) based on the quality and quantity of the premises. Of the 16 predictions made on the basis of the atmosphere hypothesis, 80% were *particular* conclusions. The results of the present study indicate that atmosphere predictions are, to some undetermined extent, confounded with the effect of the logical definition of "some."

It is important to note that there was no interaction between formal and affective terms on either time or errors. In terms of Osgood's (1963) model of the language process, affective responses are associated with relatively stable interactions based upon past associations (of the stimuli involved) with similar or dissimilar objects or events. The results of the present study attest to the relative stability of affective interactions. The relatively large amount of variance accounted for by formal terms is also consistent with Osgood's statement that the effects of quantifiers should follow a multiplicative function.

More time was spent on incompatible stimulus combinations on the posttest—a reversal of pretest results. If conceptual conflict, as outlined by Berlyne (1960), consistently functioned to increase decision time, one would not expect such a pre-posttest interaction. When *S*s have become sensitized to verbal stimuli, which presumably occurred to some extent on the pretest, time scores tend to reflect efforts to overcome negative transfer effects. A theory of conceptual conflict might usefully take into account such interactions between stimulus familiarity and response bias when complex relations are involved.

The present experiment provides for a comparison of two hypotheses: (*a*) that stimulus compatibilities increase acceptance of conclusions, and (*b*) that stimulus compatibilities facilitate reasoning. Because of the large number of invalid conclusions, an increase in the acceptance of conclusions would mean an increase in the number of errors, but fewest errors were made on conclusions with highly compatible terms.

It was noted that there was no significant difference in overall errors made on the posttest between the group which received relevant training and the group which did not. The trained group did improve significantly in dealing with *particular* propositions, and the results of their training showed up in a later test. It is apparent that overall test results may not reflect improvement in reasoning which a precise instrument can reveal. In syllogistic reasoning, improvement in understanding the logical meaning of "some" foreshadows enhanced general performance.

Results clearly revealed the operation of two response systems during reasoning; responses controlled by formal structural terms and responses controlled by affective terms. The former were relatively sensitive to training and exhibited gross changes. The latter were less influenced by training and exhibited moderate response characteristics. The formal syllogism, which can easily be extended from the standard form employed in this study, might prove a useful paradigm for studying the effects of several variables upon the two response systems in verbal problem-solving.

REFERENCES

BERLYNE, D. E. *Conflict, arousal, and curiosity*. New York: McGraw-Hill, 1960.

HAYS, W. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.

JANIS, I. L., & Frick, F. The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology*, 1943, **33**, 73–77.

KEARNS, J., ODELL, S. J., & ZARTMAN, J. A programmed introduction to logic, 1964, University of Illinois, Training Research Laboratory Contract 1834(36), Office of Naval Research.

LEFFORD, A. The influence of emotional subject matter on logical reasoning. *Journal of General Psychology*, 1946, **34**, 127–151.

MERRILL, M. D. Correction and review on successive parts in learning a hierarchical task. *Journal of Educational Psychology*, 1965, **56**, 225–234.

MOWRER, O. H. The psychologist looks at language. *American Psychologist*, 1954, **9**, 660–694.

MOWRER, O. H. *Learning theory and the symbolic processes*. New York: Wiley, 1960.

OSGOOD, C. E. On understanding and creating sentences. *American Psychologist*, 1963, **18**, 735–751.

OSGOOD, C. E., ARCHER, W. K., & MIRON, M. S. *Cross cultural generality of meaning systems*. Urbana: Institute of Communication Research, 1962.

OSGOOD, C. E., & TANNENBAUM, P. H. The principle of congruity in the prediction of attitude change. *Psychological Review*, 1955, **62**, 42–55.

SELLS, S. B. The atmosphere effect: An experimental study of reasoning. *Archives of Psychology*, New York, 1936, No. 200. Pp. 72.

SOLARZ, A. K. Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *Journal of Experimental Psychology*, 1960, **59**, 239–245.

STAATS, A. W. Meaning and word associations: Separate processes. Technical Report No. 12, 1959, Contract NONR 2794-(02), Office of Naval Research.

WILKINS, M. C. The effect of changed material on ability to do formal syllogistic reasoning. *Archives of Psychology*, New York, 1928, No. 102. Pp. 83.

(Received January 24, 1966)

Beginning with the 1967 volume, this Journal will consider manuscripts for publication as monographs. Monograph Supplements will appear under separate covers, but will be mailed with issues of the Journal to all subscribers.

Monographic manuscripts will be judged by the same standards of quality as articles in the Journal. In addition, they are expected to present comprehensive experimental investigations, frequently reporting a series of experiments, or to provide more extensive theoretical integration of a topical area than is possible in the space available to articles. Authors of monographs whose content is germane to the area of this Journal are invited to submit manuscripts to the editor.

Journal of Educational Psychology
1966, Vol. 57, No. 5, 246–252

DELAY OF GRATIFICATION AS A FUNCTION OF CHARACTERISTICS OF SOCIAL AGENTS[1]

MARCIA J. LAWTON
University of Nebraska College of Medicine

Following an irrelevant task 171 1st and 2nd graders were individually offered, by a male or female stranger or a female student teacher, a choice between an immediate, less preferred reward and a delayed, preferred one. Data on such variables as age, IQ, ability to estimate time, self-control, and attitudes toward promise-keeping were investigated in relationship to such characteristics of the social agent offering the choice as sex of the examiner, experience with the examiner, and the form in which the promise was given. The most important finding was that unfamiliar women elicited significantly ($p = .10$) fewer delayed responses than unfamiliar men or familiar women, a finding that was explained in terms of trust.

By the time a child is of school age, he is ready to be taught many new skills by significant adults other than his parents; that is, his teachers. Erikson (1959) says

> good parents feel a need to make their children trust their teachers, and therefore to have teachers who can be trusted . . . [thus] leading to the development and the maintenance in children of a *sense of industry* and of a positive identification with these who *know* things and know how to *do* things [p. 87].

The question arises as to what components are involved in a teacher-pupil trusting relationship.

The basic tenet of Mischel's delay-of-gratification approach, which consists of offering children a choice between an immediate, less preferred reward and a delayed, preferred one, is that the choice of a reward is a function of the expectancy of its occurrence and the value of the reinforcement (Rotter, 1954). Initially, studies focused on the hypothesized decline in the *value* of a delayed reward and the length of time interval, but in 1956 Mahrer and in 1958 Mischel shifted the emphasis to an analysis of trust, operationally defined as the *expectancy* for a reinforcement to follow as a function of making a particular choice with a particular social agent in a given situation. Empirically, research in this area has attempted to clarify variables relevant to preferences for immediate reward or delayed reward. Mostly using children from other cultures as subjects (*S*s), it has been demonstrated that choice of a delayed reward is related positively to intelligence (Melikian, 1959; Mischel & Metzner, 1962), to the presence of the father (Mischel, 1958, 1961b), to the need for achievement and social responsibility (Mischel, 1961a, 1961c) and negatively to acquiescence (Mischel, 1961a). Finally, research has shown no sex differences in preference for delayed reinforcement (Mischel, 1961b), but it has shown that preference for delayed reinforcement apparently decreases as the length of time increases (Mischel & Metzner, 1962), and increases with age up to a certain level depending on the particular values used (Melikian, 1959).

Specifically, the present study attempts to demonstrate the importance of characteristics of the particular social agent including the sex of the examiner, experience with the examiner, and the form in which the promise is given.

---

[1] Based on a doctoral dissertation submitted to Northwestern University, June 1963, and written under the helpful direction of Lee B. Sechrest. The material on which this article is based was obtained from children in District 65 of the Evanston School System, which was extremely cooperative in assisting the author.

Sex of Examiner

Several studies (Alden & Benton, 1951; Holtzman, 1952) have stressed the influence of the sex of the examiner in the testing situation. Also Mischel[2] found male children showed significantly greater delay preference with hypothetical delay choices when administered by a male experimenter. Inasmuch as young children have considerably more experience with women, it is possible that they build up differential expectancies of promise-keeping for men and women.

Experience with Examiner

The expectancies which a child has for promise-keeping by perfect strangers may differ in important ways from those he has towards people he knows and with whom he has had greater contact. In this study promises were made by known student teachers and unknown graduate students.

Form of Promise

Two forms, a paper note and an oral promise, are compared. It is assumed that if a child is given a secondary reinforcer in the form of a paper promise, this will be more effective than a verbal promise both because of the explicitness of the "contract" and because of the similarity to other secondary reinforcers which help to bridge delays (Renner, 1964).

Additional data were obtained for a number of characteristics thought on an a priori basis to be related: age, IQ, grade placement, attitudes toward promise-keeping, ability to control behavior, and ability to estimate time.

METHOD

Subjects

A total of 171 *S*s from five first grades and four second grades were employed in this study. The schools tended to be in a middle- to high-middle-class socioeconomic area and the overall IQ was 110. Of these children 124 came from intact homes while 18 came from broken or adoptive homes and 29 came from homes whose structure was unknown to the teachers filling out the forms.

Examiners

A total of 14 examiners were used who fell into three categories: four male strangers, four female strangers, and six female student teachers. The strangers were all graduate students in their early or mid-twenties enrolled in a clinical psychology training program. They were selected largely in terms of availability of time and willingness to participate.

Form of Promise

As a reward each *S* was offered a *choice* between two small balloons today (ImR) and one large balloon in a week (DelR). (See below under *Choice* for further explanation). The DelR was offered to half of the *S*s orally and to the other half by means of a paper note on which his name and the date of delivery were written.

Assignment

The 171 *S*s were randomly assigned to one of the following groups:

1. Male Stranger Oral
2. Male Stranger Paper
3. Female Stranger Oral
4. Female Stranger Paper
5. Female Student Teacher Oral
6. Female Student Teacher Paper

In general, each examiner saw about 10 *S*s, although one saw as few as seven and one as many as 18, divided about equally by sex, grade, and form of promise. The mean and range in ages of *S*s seen by each "stranger" examiner closely approximated the overall mean age of 84.8 months and the overall range of 72 to 102 months, but student teachers were limited to testing *S*s within their own classroom where the age range was somewhat narrower.

---

[2] W. Mischel, personal communication, October, 1962.

Procedure

Each *S* was seen individually for about 15 minutes by the examiner to which he was assigned. Because of the distance to the testing room relative to most of the classrooms, it was necessary for the examiner to come to the classroom, meet the child there, and walk to the testing room with him, chatting in a neutral way.

The administration of the actual testing involved five parts:

1. Introduction. The "stranger" examiner introduced himself as being someone (no names were used) interested in learning what TV programs children of this age most enjoy. The student teacher examiners explained they were helping a person from the university interested in learning what TV programs children of this age most enjoy.

2. Task Questions. The examiner then asked *S* three questions about his TV habits which were really irrelevant to the dimension under investigation, delay of gratification, but which served as a reason for offering *S* a choice of reward.

3. Choice Situation. Next the examiner offered *S* a reward for his assistance; namely, either two small balloons today (ImR) *or* one large balloon in a week (DelR), half with and half without the paper note. School regulations necessitated changing the candy reinforcements such as Mischel used to balloons, which Gray[3] had found effective. A pilot study, based on individual testing of 20 second graders, revealed that a choice of two small balloons immediately, versus one large balloon in a week, met Mischel's criteria. That is, in a straight choice situation the larger balloon was chosen 100% of the time whereas under a delayed condition of 1 week half chose two small balloons immediately.

In the testing situation *S*'s choice was recorded (DelR or ImR). At this point *S* was also asked to indicate the number of balloons which would have to be added (if a DelR was originally chosen) or deleted (if an ImR was originally chosen) for him to change his choice. For example, if *S* chose the large balloon, the examiner would add one, then two, then three, etc. small balloons to the two balloons already there until he said he would take the pile of small balloons instead of the large one. The number of balloons added was recorded (change-over score) and served as a measure of the strength of *S*'s initial choice; that is, the more balloons which had to be added, the stronger his original desire for the big balloon was assumed to be. It was made clear to *S* that his original choice was the balloon or balloons that he would receive and that the later choices were only hypothetical ones.

4. Promise Questionnaire. Following his making a choice, *S* was asked to answer four questions involving what it means to keep a promise and whether, in fact, he believed that people tend to keep promises. Of these four, two were answered clearly enough that they could be used for statistical comparisons. One plus or minus score was given for whether or not he believed others kept their promises and one for whether or not he mentioned some form of punishment coming to people who do not keep their promises.

5. Time Questionnaire. This set of 10 questions involved *S*'s ability to estimate time, both short-term and long-term. For example: How long before you get out of school today and how long will it be before you are married? Each question was scored on a predetermined three-point rating scale ($0 =$ substantial error, $1 =$ some error, $2 =$ substantial accuracy) of accuracy. On the basis of the overall distribution the accuracy scores were then grouped into high (11 or above), middle (6–10), or low (0–5) accuracy scores which were used for computational purposes. The communicability of the scoring system employed was checked on a random sample of 40 *S*s, for which a Pearson product-moment correlation of .99 was obtained between two scorers. The split-half reliability of the questionnaire corrected by the Spearman-Brown prediction formula was .75.

After the testing was completed, the child was returned to his classroom where he deposited his two small balloons (if ImR was chosen) or paper note (if written DelR was chosen) in his locker outside the classroom. If an oral DelR was chosen, the child received his balloon 1 week later in a group from his teacher.

6. Other Measures. Each teacher was requested to furnish control ratings and background information. Because of the paucity of material available on many questions it was deemed advisable to use only the following:

Control ratings. The teacher's ratings on a five-point scale of *S*'s ability to control impulses and delay gratification were totaled across six items for each child. These included his ability to wait his turn, to share, to ask for something, etc. The possible range of scores was from a low of six to a high of 30. A split-half reliability coefficient of .92, corrected by the Spearman-Brown prediction formula, was obtained. There were no appreciable differences between the mean control ratings given by the nine teachers, as most of them hovered around an overall mean of 20 with the range being from 17.1 to 25.6.

---

[3] Susan Gray, personal communication, November, 1962.

Father's occupation. The teacher indicated the father's occupation in almost all the cases. In three cases it was necessary to rely on the teacher's gross estimate of the socioeconomic class. Otherwise the Coleman (1959) socioeconomic class rating scale was applied to the occupational information; there was interscorer reliability on a random sample of 40 *S*s of .90.

IQ. In many of the classes quantitative IQ estimates based on previous administration of formal intelligence tests were available. However, for 68 *S*s there was no score available; in 45 of these cases the teachers furnished estimates (low, low-average, average, high-average, and superior) of the child's ability. Analyses involving IQs were done twice, once on the basis of the reported quantitative scores alone and once on the basis of all the estimates combined.

Siblings. On the basis of the information supplied by the teacher, *S* was placed in one of five categories indicating the number of siblings he had. There were none, one, two, three, and four or more.

RESULTS

Chi-squares were computed comparing the number of immediate and delayed reward choices made under the various conditions. The results for examiners in a particular group were pooled and analyses were computed on this pooled data because preliminary tests involving comparisons between the various individual examiners within each of the groups were all nonsignificant (four male strangers, $\chi^2 = 4.50$, $df = 3$, $p < .25$; ten female examiners, $\chi^2 = 8.27$, $df = 9$, $p < .30$; six known examiners, $\chi^2 = 8.94$, $df = 5$, $p < .15$; eight unknown examiners, $\chi^2 = 8.64$, $df = 7$, $p < .30$).

An additional analysis was performed on changeover scores for various experimental groups in cases of children choosing DelR. The analysis of variance used on this measure involved a range of changeover scores from one to five. Specific results follow.

TABLE 1

COMPARATIVE NUMBER OF CHOICES ACCORDING TO SEX AND EXPERIENCE WITH EXAMINER

Examiner	Immediate	Delay
Male stranger	17	44
Female stranger	26	30
Female student teacher	15	39

Variable I (Sex of the Examiner)

In analyzing the role played by the sex of the examiner, it was found that the chi-square (1.56) based on the number of immediate versus delayed choices (see Table 1) was not significant for male versus pooled female examiner groups (stranger and known). However, when the results were based on comparisons between male strangers and female strangers, the chi-square obtained (4.30) was significant ($p < .05$). The male strangers elicited significantly more delayed choices than did the female strangers. Further analysis of the data indicates

TABLE 2

CHANGEOVER SCORES FROM DELAYED TO IMMEDIATE CHOICE

*S*s	Type of promise					
	Oral			Paper		
	N	$\overline{X}$	*SD*	*N*	$\overline{X}$	*SD*
By sex						
Male	19	4.10	1.62	25	3.64	1.60
Female	38	3.34	1.79	31	4.84	.58
By experience with examiner						
Known	20	4.30	1.15	19	4.95	.23
Unknown	37	3.21	1.86	37	3.97	.47

that there was no interaction effect between the sex of the examiner and the sex of the subject ($\chi^2 = 2.87$).

The top portion of Table 2 contains changeover scores for the two sexes of examiners. An analysis of variance yielded nonsignificant F values ($F = .20$, $df = 1$).

Variable II (Experience with the Examiner)

Both choice scores and changeover scores for female strangers versus female student teachers produced significant differences. Examiners who were known to the children tended to elicit significantly more delayed choices than did examiners with whom the children had had no experience, as shown in Table 1 ($\chi^2 = 4.04$, $df = 1$, $p < .25$). Furthermore, once the child decided to wait, more incentive (small balloons) was required with known examiners than with unknown examiners, as shown in Table 2 ($F = 4.76$, $df = 1$, $p < .05$).

Variables I and II Combined

The combined efforts of these two variables, sex of the examiner and experience with the examiner, were analyzed with respect to number of choices and changeover scores. A chi-square of 4.76 was obtained between the three groups (male strangers, female strangers, and female student teachers) which was just significant at the .10 level of confidence (see Table 1). In essence, children chose 2½ times as many delayed rewards as immediate ones when they were offered by either the male strangers (44 to 17) or the student teachers (39 to 15). However, when the offer came from the female strangers, there was a greater reluctance to choose delayed rewards, the split being about as many immediate rewards as delayed ones (26 immediate, 30 delayed). Thus, it was the female strangers who formed the deviant group.

TABLE 3

CHANGEOVER SCORES FROM DELAYED TO IMMEDIATE CHOICE ACCORDING TO SEX OF SUBJECT

Group	Male S			Female S		
	N	$\bar{X}$	SD	N	$\bar{X}$	SD
Male stranger	22	3.82	1.53	22	3.82	1.61
Female stranger	15	3.40	1.91	15	3.07	.60
Female student teacher	18	4.61	.99	21	4.62	.87

When an analysis of variance was performed on the changeover scores obtained by the three groups a significant F of 6.36 (beyond .01 level of confidence) was derived (see Table 3). A Duncan multiple-range test was applied to the data and it revealed that each of the three groups differed significantly from the others. Thus, female student teachers obtained significantly higher changeover scores ($M = 9.23$) than did male strangers ($M = 7.64$) who, in turn, obtained significantly higher ones than did female strangers (6.47). Therefore, it appears that Ss were most willing to wait with student teachers, less willing to wait with male strangers, and least willing to wait with female strangers.

Variable III (Form of Promise)

The difference between the number of immediate and delayed rewards obtained under the oral promise condition and the paper promise condition was nonsignificant when computed for χ^2 ($\chi^2 = .002$). In fact, the exact number of immediate and delayed rewards chosen under the oral (29 and 57) and paper (29 and 56) conditions was almost identical. Similarly, a separate analysis for the six combinations for examiner and form of promise was nonsignificant (see Table 4; χ^2

$= 7.33$, $df = 5$, $p < .20$). An analysis of variance of changeover scores (see Table 2) yielded nonsignificant F's for the form of promise when analyzed across experience with the examiners ($F = 0.00$).

There was a significant interaction effect on changeover scores for sex of examiner and form of promise (see Table 2; $F = 3.81$, $p < .05$). Observation of the data indicates that male examiners tended to elicit higher changeover scores under the oral condition than did female examiners, whereas female examiners tended to elicit much higher changeover scores under the paper condition than did the male examiners. Thus, for female examiners alone, a paper note strengthens the children's willingness to wait for the delayed reward.

Analysis by chi-square of other variables (attitude toward promise keeping, time estimation, ratings of self-control, IQ, age, and grade placement) revealed no significant differences except that first graders differed ($\chi^2 = 3.97$; $p < .05$) on the type of choice made when it was offered by male strangers and by female strangers. They were more willing to choose the delayed reward when it was offered by a male stranger than when by a female stranger. This finding did not hold for second graders. There was no significant difference in the choices obtained by female strangers as contrasted with female student teachers when compared separately by grades.

TABLE 4

NUMBER OF IMMEDIATE AND DELAY CHOICES IN EACH EXPERIMENTAL CONDITION

Form of promise and examiner	Immediate	Delay
Male oral	10	19
Female oral	13	18
Student teacher oral	6	20
Male paper	7	25
Female paper	13	12
Student teacher paper	9	19

DISCUSSION

Results of this investigation show that the probability of an S choosing an immediate or a delayed gratification is a function of the characteristics of the person offering the choice and making the promise. The author prefers to view the response of the child in terms of "trust," but there are certainly alternatives. When S is placed in situations resembling the immediate-delayed reward choice, his response will reflect, among other things, the expectation he has that the promised larger reward will, in fact, be delivered. That aspect of such situations can be separated conceptually from S's ability or inability to evaluate the period of delay, the relative worths of the rewards offered, etc. It is just that expectancy of fulfillment that defines trust as a response.

The most important finding here is that unfamiliar women are trusted less than either familiar women or unfamiliar men. Of course it must be recognized that "familiarity" is, in this investigation, a rather gross description that is confounded with other variables such as teacher-nonteacher status and the like. Nonetheless, it appears to the writer that the strange-familiar dimension must be the most salient one for first- and second-grade children taken from their classrooms by an adult. It does not seem surprising that children should be more willing to risk delay of gratification more with their own teacher than with a strange female. What is more difficult to comprehend is why delay is more acceptable when the agent is a strange male. At present we can do no more than speculate, but two ideas that present themselves are that: (*a*) Males are generally more prestigeful in an elementary school situation and, hence, may be more trustworthy, and (*b*) the nature of a child's interactions with male and female figures almost insures that the female will have had more opportunity to break promises, for example, mothers must, perforce, break what were to the child considered ironclad agreements.

One additional—and intriguing—possibility is that for most children the choice of an immediate reward represents a self-indulgence which they are not always permitted. That is, we may put the emphasis on the gratification of an immediate reward rather than upon the somewhat puritanical promise of pleasures to come. In that case, we could ask what kinds of persons are likely to elicit a self-indulgent and perhaps immature response from a child. The writer does not find it a strain to suppose that the warm and friendly strange female graduate students employed in this investigation would foster more self-indulgence in a child than either a teacher or a strange male. More research is called for on this point.

The failure in this investigation to find a relationship between the age and IQ measures and reward choice is not consistent with the findings of other investigators (Melikian, 1959; Mischel & Metzner, 1962), but the sample used here was quite homogeneous with respect to age and probably with respect to intelligence also.

The "paper promise" proved no more effective than the verbal promise in eliciting delay choices, a finding which might be explainable either in terms of the unfamiliarity of the children with promissory notes or in terms of a possibility that trust was not really so much at issue. It is of interest that the use of the note increased delay choices when the agent was a strange female, but that finding will need elucidation in other research before it becomes clearly interpretable. It makes sense, however, as a manifestation of bolstering trust.

REFERENCES

ALDEN, P., & BENTON, A. L. Relationship of sex of examiner to incidences of Rorschach responses with sexual content. *Journal of Projective Techniques,* 1951, **15,** 230–234.

COLEMAN, R. P. Social class in Dawson City. Unpublished doctoral dissertation, Committee on Human Development, University of Chicago, 1959.

ERIKSON, E. Identity and the life cycle. *Psychological Issues,* 1959, **1,** 1–171.

HOLTZMAN, W. H. The examiner as a variable in the Draw-A-Person Test. *Journal of Consulting Psychology,* 1952, **16,** 145–148.

MAHRER, A. R. The role of expectancy in delayed reinforcement. *Journal of Experimental Psychology,* 1956, **52,** 101–105.

MELIKIAN, L. Preference for delayed reinforcement: An experimental study among Palestinian Arab refugee children. *Journal of Social Psychology,* 1959, **50,** 81–86.

MISCHEL, W. Preference for delayed reinforcement: An experimental study of cultural observation. *Journal of Abnormal and Social Psychology,* 1958, **56,** 57–61.

MISCHEL, W. Delay of gratification, need for achievement, and acquiescence in another culture. *Journal of Abnormal and Social Psychology,* 1961, **62,** 543–552. (a)

MISCHEL, W. Father-absence and delay of gratification: Cross-cultural comparisons. *Journal of Abnormal and Social Psychology,* 1961, **63,** 116–124. (b)

MISCHEL, W. Preference for delayed reinforcement and social responsibility. *Journal of Abnormal and Social Psychology,* 1961, **62,** 1–7. (c)

MISCHEL, W., & METZNER, R. Preference for delayed reward as a function of age, intelligence, and length of delay interval. *Journal of Abnormal and Social Psychology,* 1962, **64,** 425–432.

RENNER, K. E. Delay of reinforcement: A historical review. *Psychological Bulletin,* 1964, **61,** 341–361.

ROTTER, J. B. *Social learning and clinical psychology.* New York: Prentice Hall, 1954.

(Received April 26, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 5, 253-270

REFINEMENT AND TEST OF THE THEORY OF FLUID AND CRYSTALLIZED GENERAL INTELLIGENCES

JOHN L. HORN
University of Denver

AND

RAYMOND B. CATTELL
University of Illinois

The 23 factors previously identified as representing primary mental abilities and 8 factors previously defined as general personality dimensions were factored, using a sample of 297 adults, to provide evidence for hypotheses stipulating that general visualization, fluency, and speediness functions, as well as fluid and crystallized intelligence functions, are involved in the performances commonly said to indicate intelligence. 9 principal axes factors were sufficient to account for the observed, generally positive, intercorrelations among the 31 primary factors. These were rotated blindly to oblique simple structure. The resulting structure was consistent with predictions based upon refinements of the general theory of fluid and crystallized intelligence. Positive manifold for the intercorrelations among the 2nd-order factors was interpreted as indicating a social fact of interdependence between intraperson and environmental influences determining behavioral attributes.

The theory of fluid and crystallized intelligence (Cattell, 1957, 1963; Horn, 1965b; Horn & Cattell, 1966a, 1966b) seriously questions the notion that there is a unitary structure which can be designated general intelligence. It questions the belief—often implicit, but expressed clearly in a recent article by McNemar (1964)—that the conglomerate measured by combining subscores from a collection of intellectual tests is the best estimate of intelligence. But the theory also questions the contention that it is theoretically and practically most worthwhile (at this stage of development in this field) to "splinter" the intellectual domain into a very large number of narrow, slightly distinct abilities, as in the recent work by Guilford (1959, 1966) and his coworkers (Guilford & Merrifield, 1960). The theory takes issue with the popular belief that general verbal and quantitative abilities are truly general, that is, represent distinct, unitary concepts at a high level of organization of all human abilities. And although the theory is in some respects similar to theories stipulating a distinction between general verbal-education (V:ed) and general practical-mechanical (k:m) abilities (Burt, 1949; Vernon, 1950) it differs from these conceptions with respect to both: (*a*) specification of the abilities which go into the major dimensions and (*b*) the development of abilities in childhood and the adult years. Looked at specifically in terms of the concerns of this study, the theory of fluid and crystallized intelligence contains clear-cut statements about the expected higher-order factorial structure among *primary mental abilities*, that is, abilities of the type defined initially in Thurstone's (1938) pioneering researches and subsequently in many other investigations using similar methods (summarized by French, 1951; French, Ekstrom, & Price, 1963; Guilford, 1959, 1966; Guilford & Merrifield, 1960).

The fluid-crystallized theory argues that the primary abilities which can be said to involve intelligence to any considerable degree are organized at a general level into two principal classes or dimensions. One of these, referred to as *fluid intelligence* (abbreviated Gf) is said to be the major measurable outcome of the influence of biological factors on intellectual development—that is, heredity, injury to the central nervous system (CNS) or to basic sensory structures, etc. The other broad dimension, designated *crystallized intelligence* (abbreviated Gc), is said to be the principal manifestation of a unitariness in the influence of experiential-educative-acculturation influences. Each dimension is, according to theory, so pervasive relative to other ability structures and so obviously of an intellectual nature that each deserves the name "intelligence."

Studies having particular relevance for the present research may be grouped into three classes: (*a*) factor analytic investigations at what may be regarded as the second (or higher) order among primary ability factors, (*b*) factor analyses of the Wechsler scales, and (*c*) results from work employing hierarchical methods for defining dimensions among ability performances. Horn (1965b) has reviewed the findings from studies in these various areas and has pointed out that these imply refinement of the general theory. Some of the principal points of refinement, and thus some of the principal hypotheses of the present study, may be summarized as follows:

Gf and Gc should be seen as only two among several general influences manifested in performance on putative tests of intelligence. Among these other influences are:

1. General visualization (Gv), a process involving visual acuity, sensitivity to light, breadth and width of visual field, depth perception, etc., and manifested in speeded performances of visualizing the movements and transformations of spatial patterns, maintaining orientation with respect to objects in space, unifying disparate elements and locating a given configuration in a visual field. The hypothesis implied is that Gv will be strongly manifested through primary factors like Vz (Visualization), Cf (Flexibility of Closure), Cs (Speed of Closure), DFT (Figural Adaptive Flexibility) and S (Spatial Orientation), but that it will be involved to a lesser degree in the other primaries involving spatial materials, such as CFR (Figural Relations), CFC (Figural Classifications) and Mk (Mechanical Knowledge).

2. General fluency (F), a process reflected in tasks indicating facility in recalling and recognizing labels for cultural concepts irrespective of the subtlety of understanding of these. The process could relate to the size of store of language elements and/or to factors determining quickness in bringing such elements from long-term memory into immediate awareness. In any case it is to be supposed that the process will be manifested strongly in those primaries—like Fa (Associated Fluency), Fi (Ideational Fluency) and Fw (Word Fluency)—which require speedy production of concept labels in a loosely structured area of meaning, but it will appear, if more faintly, in primaries like V (Verbal Comprehension) and EMS (Experiential Evaluation), since these require some facility in recall of concept labels although the principal requirement of the tasks is not this.

3. General Speediness (Gs), a quickness of performance involved to some extent in most speeded intellectual tasks but independent of the above-mentioned functions and not indicative of the capacity to perceive

relations. This quickness is expected to show up most clearly in very simple tasks, such as those of the primaries P (Perceptual Speed) and N (Number Facility), but the function will be related, also, to the attitudes, self-evaluations, etc., found to define UI-16 (Assertive Ego) and UI-21 (Exuberance).

In the same space as the above dimensions, fluid intelligence will represent processes of reasoning in the immediate situation in tasks requiring abstracting, concept formation and attainment, and the perception and eduction of relations. It will be measured most purely when task materials are culture fair; that is, the fundaments are either novel for all persons being measured or else are extremely common, overlearned elements of a culture, and the aids (Cattell, 1963; Horn, 1965b) needed to attain solutions are not those made available, by favored educational opportunity, to some persons and not to others among those measured. Thus it can be predicted that Gf will be manifested in the reasoning primaries CFR, CFC, I (Induction), CMR (Semantic Relations), R (General Reasoning), and Rs (Formal Reasoning) and it will appear in primaries requiring the subject to retain elements in immediate awareness, such as Ms (Memory Span), Ma (Associative Memory), and the simple Clerical Speed primaries, P and N.

At the same level of generality as Gf and involving an equally broad set of intellectual tasks, the crystallized intelligence function will indicate breadth of awareness and subtlety of relations previously perceived, concepts previously attained, etc., as indicated in tasks requiring recognition or recall of such relations, but it will also represent ability to reason in the immediate situation in tasks requiring abstracting, concept formation and attainment, perception, and eduction of relations. Thus, in contrast to Gf, Gc will be measured most purely under conditions in which the subject must use concepts and aids representing relatively abstruse elements of the collective intelligence of a culture. Therefore Gc will be manifested in "recall" primaries, like V (Verbal Comprehension) and Mk (Mechanical Knowledge), and also in reasoning primaries, like CMR, CMC, R, Rs, J (Judgment), and EMS (Experiential Evaluation).

This development of the theory introduces a notion of alternative mechanisms in the performances that are commonly assumed to represent the operation of intelligence. That is, many intellectual tasks allow one to employ either fluid intelligence or crystallized intelligence to arrive at a correct answer. For example, in a problem like this:

> In a hospital containing 100 patients some are one-legged but wearing shoes and half of the remainder are barefoot. How many shoes are being worn?

a person may use an elementary induction process largely representative of fluid intelligence or he may use the techniques of formal algebra, these being part of the collective intelligence of our culture and thus of Gc. A primary factor which involves tasks of this type, such as R, may therefore be expected to relate about as much to Gf as to Gc. On this basis, then, it can be predicted that the primaries R, CMR, CMC, Rs, and J will help to define both fluid and crystallized general intelligence, as suggested previously. On this basis, too, it can be predicted that primaries allowing for use of both visualization and fluid intelligence processes, like CFR and CFC, will be involved in the factorial identification of both Gv and Gf. Similarly, the fluency primaries may be expected to relate to both Gc and F.

In the only previous study concerned specifically with the concepts of fluid and crystallized intelligence, Cattell (1963) found two second-order dimensions which could be interpreted in accordance with hypotheses about the processes of Gf and Gc. However, the sample of primaries in this study was of very limited range and not such that it would permit test of several of the hypotheses outlined or implied above. For example, the factor Cattell identified as Gf, involving CFR, CFC, DFT, and S, would seem (according to the above refinements) to be confounded with Gv; the factor he interpreted as Gc, defined by V, N, and I was narrow relative to the general hypothesis stated above; the hypothesis of alternative mechanisms could not be examined and no Gs or F factor was indicated. It is of some interest therefore, to see if the theory can be generalized in the way suggested here.

More than this, it is of some interest to explore the second order factorial structure of a broad sample of primary ability factors whether or not the Gf-Gc theory is taken as the motivation for this. For although there have been several factorial studies at what can be regarded as the second order, as reviewed by Horn (1965b), none of these employed a representative sample of primary factors presently known and replicated, none was designed specifically to reveal intricacies of higher order organization (with the result that adequate hyperplane material was usually lacking), and in most cases the primaries themselves were rather narrowly defined.

The major purposes of the present study are thus indicated: to illustrate the structure in a comprehensive sample of primary mental ability factors with the aim of determining whether or not this is generally consistent with the process descriptions outlined in the above refinement of the theory of fluid and crystallized intelligence.

OPERATIONAL DEFINITIONS OF CRUCIAL VARIABLES

Horn and Cattell (1965) have recently reemphasized the desirability of using more than one test to measure and represent a primary factor. If a single test is used, the resulting measure is likely to represent the test specific factor to a considerable extent and to contain substantial variance on an unwanted vehicle factor. To heed this advice presents problems when the aim, as in the present study, is representatively to sample from a broad domain of factors and yet testing time must, for various reasons, be limited to no more than five hours. In this study these problems were solved, though not ideally, by using two or more tests to measure most of the primaries having pointed relation to the major hypotheses and measuring the remaining primaries with one test which previous research had shown to be a relatively pure marker for the factor.

Cattell (1963) pointed out that in order for Gf and Gc general ability dimensions to be uniquely defined by a simple structure it is quite necessary for an analysis to involve nonability primary factors—to furnish hyperplanes against which to rotate the ability factors. Consideration must be given to sampling not only from the ability domain but also from the domain of temperament and general personality factors. In the present investigation roughly one-third of the primaries included for analysis were taken from the Q-data realm or were general objective test factors believed to be large independent of the processes of intellectualization, as such.

The tests used to achieve measurement of the primary factors are listed with brief descriptions in Table 1. A name following a variable title in the table indicates that the variable was taken either from a test published under the authorship of the person whose name is listed or else from a study reported in the research literature by this person. The abbreviations for primary factors are those used by French et al. (1963) if these authors list the factor. If they do not, then the abbreviation is taken either from French (1951), Guilford and Merrifield (1960), or Horn (1965b).

In the listings defining the "Scoring," R means number right, W means number wrong, N means number of responses given in specified time on a speeded open-ended test, K means score determined according to a specific key, and the other symbols merely indicate the way in which the variable was obtained from these basic kinds of scores. Thus 20 − W, for example, means that the number of wrong answers given in a test was subtracted from the constant 20 in order to obtain a variable measure in which a high score would indicate "good" performance (i.e., few wrong).

Most of the reliability estimates given in Table 1 were based upon separately-timed half-tests. This was true in all cases where a test was speeded. Total work time does not include the time required to give instructions.

Test scores were combined in the manner indicated to obtain the primary factor scores that were actually analyzed. In each case the test variable was converted to standard score form and added in with unity nominal weight.

The symbols of the column to the far right in Table 1 indicate the major second-order hypotheses for each of the primary factor variables. The order of listing of these symbols in a particular row indicates primacy—that is, the first-listed is considered to be the major hypothesis; the second indicates a lesser hypothesis. PSI in this column stands for a second-order positive self-image dimension found among nonintellectual primaries (Horn, 1963). The other symbols were defined above.

It is likely that most investigators will accept this sampling of primaries as reasonably representative of what is now known about first-order factors. Of course, there are other fairly well-established primaries which could, with justification, have been included in this study. Nevertheless the sample here employed is at least somewhat stratified to cover the area of cognitive abilities and, in any case, is the most extensive and representative yet achieved in this kind of study.

METHOD

Subjects

The general theory argues that the influences which produce the distinction between Gf and Gc operate through development. As development proceeds, the Gf-Gc distinction becomes clearer. A detailed presentation of the facts and theoretical considerations which support this position has been provided elsewhere (Horn, 1965b). In general, however, this conclusion follows from the fact that although educative and nutritive influences are to some extent interdependent and mutually supportive, there is no *necessary* relationship between

TABLE 1

Primary Factor Variables and Tests Used for Estimation

Matrix number	Primary title	Primary symbol	Test title and source	r_{xx}	Total work time (seconds)	Number of items	Scoring	Hypotheses
1	Induction	I	Letter Grouping (Botzum)	67	240	24	R	Gf, Gs
			Number Series (Botzum)	59	244	20	R	
2	Intellectual Speed	SP	Test A(2) Series (Furneaux)	63	200	20	lnR	Gf, Gs
3	Carefulness	C	Test B(1) Series (Furneaux)	57	200	20	20 − W	Gs(−)
			Figure Classify (Cattell III)	45	240	14	20 − W	
			Practical Estimates (Guilford)		420	20	20 − W	
			Subtracting (Horn)	75	60	20	9 − W	
			Dividing (Horn)	66	60	20	20 − W	
			Fractions-Decimals (Horn)		180	20	20 − W	
4	Intellectual Level	L	Test B(2) Series (Furneaux)	72	900	35	% R	Gf
5	Figural Relations	CFR	Figure Series (Cattell II, III)	69	360	24	R	Gf, Gv
			Topology (Cattell II, III)	65	300	16	R	
			Matrices Speed (Cattell II)	56	180	12	R	
			Matrices Power (Cattell III)	53	360	13	R	
			Figure Classify (Cattell II)	52	240	14	R	
6	General Reasoning	R	Problem Solving (Horn)	70	600	12	R	Gf, Gc
7	Adaptive Flexibility	DFT	Match Arrangements (Guilford)	69	420	20	R	Gv, Gf
8	Spatial Orientation	S	Cards (Botzum)	86	240	54	R	Gv
			Figures (Botzum)	85	240	52	R	
9	Visualization	Vz	Form Boards (Thurstone)	81	240	28	R	Gv
10	Associative Memory	Ma	Cued Nonsense Memory (Horn)	68	280	20	R	Gf
			Cued Meaningful Memory (Horn)	87	288	26	R	
11	Semantic Relations	CMR	Common Word Analogies (Horn)	65	390	30	R	Gf, Gc
			Abstruse Word Analogies (Horn)	67	390	30	R	Gc, Gf
12	Verbal Comprehension	V	Vocabulary (Horn)	93	900	50	R	Gc
			General Information (Horn)	84	420	28	R	
13	Mechanical Knowledge	Mk	Mechanical Information (AAF)	71	150	15	R	Gc
			Tool Identification (AAF)	72	150	15	R	
14	Formal Reasoning	Rs	False Premises (Thurstone)	41	360	22	R	Gc, Gf
			Influences (Guilford)	66	480	14	R	
15	Experiential Evaluation	EMS	Social Situations (Guilford)	61	360	16	R	Gc
16	Associational Fluency	Fa	Controlled Associations (Thurstone)	73	180		N	F, Gc

Table 1—Continued

Matrix number	Primary title	Primary symbol	Test title and source	r_{xx}	Total work time (seconds)	Number of items	Scoring	Hypotheses
17	Ideational Fluency	Fi	Things Round (Taylor)		180		N	F, Gc
			Ideas (Man up the Ladder) (Taylor)		180		N	
18	Number Facility	N	Adding (Horn)	71	70	20	R	Gc, Gs, Gf
			Multiplying (Horn)		90	20	R	
			Mixed Operations (Horn)		150	20	R	
19	Speed of Closure	Cs	Backward Reading (Botzum)	89	165	40	R	Gv, Gs
			Street Gestalt (Botzum)	73	210	24	R	
20	Flexibility of Closure	Cf	Designs (Botzum)	85	120		R − W + 20	Gv, Gs
21	Speed of Copying	Sc	Forward Writing (Cattell)		30		N	Gs
			Forward Printing (Cattell)		30		N	
22	Writing Flexibility	Wf	Backward Writing (Cattell)	58	120		N	Gs
23	Perceptual Speed	P	Match Letters & Numbers (Cattell)	81	90	50	R	Gs
			Rapid Cancellation (Cattell)	86	90		R	
24	Assentive Ego	U16	Agree With Platitudes (Cattell)	75	120	20	K	Gs
			Highbrow Tastes (Cattell)	71	180	20	K	
25	Exuberance	U21	Experience Claimed (Cattell)	76	180	20	K	Gs
			Confidence in Skill (Cattell)	55	180	20	K	
26	Self-Sentiment	U36	Utilities Choice (Cattell)	49	150	16	K	Gc, PSI
			Autistic Distortion (Cattell)	43	150	16	K	
27	Early Risks	ER	Early Risks Claimed (Horn)	86	135	26	K	PSI
28	Unbound Anxiety	U24	Test Anxiety (Sarason)	82	240	30	K	PSI (−)
			16 PF. Q4 (Cattell)					
29	Adventurousness	H	16 PF. H (Cattell)	73	180	26	K	PSI
30	Experimenting	Q_1	16 PF. Q_1 (Cattell)	64	150	20	K	PSI
31	Sensitivity	Ip	16 PF. I (Cattell)	57	150	20	K	PSI (−)

them. Thus over time their cumulative effects can tend to independence. It follows that the distinction between Gf and Gc will be difficult to draw in early childhood, will become gradually more clear-cut as development proceeds through childhood and will become most evident in adulthood.

The general theory also argues that in viewing the Gf-Gc distinction from the point of view of individual differences study, it will be most clear-cut in a sample of people wherein a wide variety of background influences would have operated to produce presently measurable abilities. When educational opportunities are closely tied to capacity, as represented in Gf, the development of Gc will tend to be closely tied to the development of Gf and the distinction between the two will be difficult to draw in empirical analyses. But when individuals differ considerably with respect to a variety of background influences, then it is likely that in many cases educational opportunities will be premised on many factors other than the capacity represented by Gf. Hence the distinction between Gf and Gc will show up most clearly in analyses with samples wherein there is heterogeneity with respect to background variables.

These considerations led the authors to seek a sample of adults from various walks of life, a sample that was nearly as heterogeneous as "adults-in-general," if not truly representative of this population.

The tests listed in Table 1 were administered to a total of 297 persons: adults and older adolescents at the Stateville, Pontiac, and Dwight State Prisons in Illinois; the Illinois Soldiers' and Sailors' Children's School; Canon City Penitentiary in Colorado; and the unemployment rolls at the Colorado State Employment Office in Denver.[1] Of the total, 215 were males. All subjects were volunteers. They were offered information about their performances as an inducement for volunteering and for doing their best.

The age range was 14 to 61 years. However, there was only one 14-year-old, two persons aged 61, one aged 56, one 55, and one 52, the bulk of the sample thus being between 16 and 51 years of age. The mean for age was 27.6 years. The standard deviation was 10.6 years.

Test Administration Procedures

The tests were administered in two sessions, each lasting approximately 2½ hours, one in the morning and one in the afternoon. All subjects did the tests in the same order, although the break points between the morning and afternoon sessions were not precisely the same for all groups. The same person administered the tests and the same stop watch was used throughout. Every effort was made to keep the conditions of administration the same for all groups.

The subjects were very cooperative. As nearly as it could be judged from "involved observations," their motivation was generally quite good in spite of the fact that the testing sessions were rather long.

ANALYSES AND RESULTS

The 31 primary factors listed in Table 1, along with age and years of education, were intercorrelated by the product-moment formula. The results are presented in Table 2. With this size sample a correlation of about .11 or .12 is significantly different from zero at the .05 level; an r of .15 is similarly significant at the .01 level. Using Hotelling's (1940) test, a difference of about .15 between two correlations would generally be significant at the .01 level.

Factoring proceeded with the first 31 variables of Table 1—that is, excluding age and education, since these are nonbehavioral variables, logically distinct from the performances represented by the measurement operations of the other variables. The number of factors was estimated by use of what may be referred to as the Kaiser-Dickman-Guttman (KDG) rationale (see Horn, 1965a). This criterion suggested nine factors. Principal factor calculations were repeated 25 times, each time replacing the diagonal element for a variable with that variable's communality estimate from the previous calculations. Since the communalities determined on the 25th iteration differed from those determined on the 24th iteration by only trivial amounts in the third and fourth decimal places, factoring was stopped at this point. The correlation matrix reproduced from the resulting nine factors very closely approximated the original correlation matrix in all off-diagonal entries, there being no residual larger than .10. The trace in the reduced space was 16, indicating that common variance was

---

[1] For invaluable help in securing this sample of subjects we particularly thank Julia Bates, staff psychologist at the Illinois Soldiers' and Sailors' Children's Home; Arthur V. Huffman, State of Illinois Criminologist; Wilson Meeks, Chairman of the Classification Board at Stateville-Joliet Prisons; Stow E. Syman, Sociologist at Pontiac Prison; Bernard Robinson, Sociologist at Dwight Prison; George Levy, Senior Psychologist at Colorado State Penitentiary, and David J. Wilson, Denver Department of Welfare.

TABLE 2

INTERCORRELATIONS OF ABILITY AND PERSONALITY FACTORS

Matrix number	Symbol	Communalities in diagonal																														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	I	64																														
2	SP	39	39																													
3	C	12	−10	48																												
4	L	53	36	30	57																											
5	CFR	67	39	21	53	75																										
6	R	42	11	21	36	36	36																									
7	DFT	40	24	13	34	46	27	35																								
8	S	46	31	03	25	49	19	39	53																							
9	Vz	45	22	11	36	54	32	44	47	58																						
10	Ma	37	28	10	35	46	20	26	26	32	38																					
11	GR	48	18	22	49	47	42	35	21	42	34	73																				
12	V	31	09	27	40	25	39	27	10	35	20	76	78																			
13	Mk	34	05	07	18	27	38	32	30	41	12	46	45	62																		
14	Rs	45	21	20	39	45	36	29	14	26	33	60	55	30	51																	
15	EMS	21	12	06	19	17	21	15	14	13	14	32	44	20	35	32																
16	Fa	08	05	04	12	13	12	11	−04	19	24	35	42	13	25	23	57															
17	Fi	24	19	−01	24	25	13	23	13	21	26	36	34	21	28	24	53	46														
18	N	45	29	13	44	33	32	33	17	32	35	45	44	30	35	30	27	36	53													
19	Cs	42	33	05	38	55	18	32	42	42	42	33	23	16	31	21	25	26	28	52												
20	Cf	46	20	16	36	47	25	37	38	55	34	35	28	33	31	19	11	23	32	48	47											
21	Sc	22	23	01	21	22	05	19	18	22	30	17	18	03	19	25	25	41	39	29	24	58										
22	Wf	06	25	00	03	19	−07	12	14	12	17	−02	−03	−15	01	13	19	26	13	29	11	49	46									
23	P	37	40	00	37	36	15	19	36	30	36	23	22	05	22	32	16	30	52	43	27	47	32	62								
24	U16	−03	−02	11	07	−08	03	02	−12	−11	−12	09	19	−02	04	10	16	03	−06	01	−02	−02	−04	03	37							
25	U21	14	19	−02	13	22	06	18	11	23	16	18	14	04	10	05	22	21	21	31	16	23	26	28	16	35						
26	U36	14	08	08	14	03	09	12	06	09	05	26	32	24	21	24	21	09	18	10	02	05	−11	06	17	05	25					
27	ER	18	16	03	10	32	08	15	28	23	20	12	04	22	20	10	06	10	06	29	15	08	10	22	−10	26	−02	79				
28	U24	−14	−05	−04	−18	−09	−18	−19	−11	−20	−04	−37	−39	−33	−24	−14	−15	−17	−14	−10	−14	−09	−15	−04	−12	−05	−15	04	50			
29	H	07	13	−05	07	08	02	07	05	07	15	20	18	14	11	09	17	15	21	16	04	20	06	16	21	35	−01	10	−49	80		
30	Q_1	11	−03	22	19	13	22	19	08	17	04	32	42	35	23	20	21	18	17	06	20	06	−11	06	26	09	19	12	−35	21	39	
31	I_p	−10	01	11	08	−04	−07	−08	−19	−05	−02	06	14	−26	01	06	22	07	−03	05	−02	01	12	03	22	11	12	−17	14	−08	02	42
32	Educ	27	20	16	38	19	20	22	03	21	08	42	53	24	34	27	22	24	42	14	24	26	12	32	18	26	14	−10	23	19	16	14
33	Age	−12	−13	14	05	−29	10	04	−14	03	−16	22	54	26	20	29	14	05	20	−18	05	01	−11	04	24	−05	21	−16	29	11	34	10

TABLE 3

OBLIQUE (SIMPLE STRUCTURE) FACTORS

Matrix number	Symbol	Gf I	Gc II	Gv III	Gs IV	PSI V	F VI	PRM VII	ER2 VIII	C2 IX	
1	I	*55*	12	15	−03	−02	−09	−06	−05	−06	
2	Sp	*30*	−07	05	10	05	−08	07	−02	−21	
3	C	05	07	06	07	−05	−14	12	06	*62*	
4	L	*47*	12	08	04	03	−11	12	−07	18	
5	CFR	*48*	−08	*35*	−07	00	01	01	08	09	
6	R	*26*	*30*	06	−04	04	−03	−12	−02	13	
7	DFT	01	09	*42*	03	00	00	−01	05	03	
8	S	04	−04	*48*	08	−07	−20	−07	04	−09	
9	Vz	−06	07	*50*	02	−05	08	−04	−04	01	
10	Ma	*31*	−02	05	10	05	15	−09	04	02	
11	CMR	*30*	*50*	06	−13	07	*29*	02	−01	02	
12	V	08	69	−06	00	00	*26*	10	−02	09	
13	Mk	−15	*48*	15	−08	00	04	−*28*	08	−07	
14	Rs	*29*	*40*	−10	−06	00	19	−04	12	05	
15	EMS	−08	*43*	−12	23	−11	05	04	08	−02	
16	Fa	−10	*35*	−06	07	01	*60*	16	02	−10	
17	Fi	−03	*25*	02	*25*	01	*42*	−04	−03	−10	
18	N	*20*	*29*	−10	*34*	04	04	−15	−07	03	
19	Cs	18	−05	*31*	06	01	10	17	07	−06	
20	Cf	−03	03	*45*	08	−05	02	01	−08	10	
21	Sc	−05	00	−01	*63*	06	09	−10	−07	08	
22	Wf	−13	−23	06	*46*	−05	14	07	01	06	
23	P	*20*	06	−03	*48*	00	−14	04	06	−03	
24	U16	−04	18	−08	−05	15	−07	*48*	−04	08	
25	U21	−03	−08	10	10	23	12	*24*	15	−04	
26	U36	01	*43*	−02	−06	−12	06	20	−02	−13	
27	ER	00	06	−01	−02	−01	00	−14	*81*	05	
28	U24	−05	−*26*	−07	06	−*47*	−01	09	14	03	
29	H	−14	−06	−12	07	*82*	00	−01	02	−03	
30	Q_1	−18	*39*	−15	02	14	−01	09	08	20	
31	Ip	04	04	−02	−04	−15	20	*54*	−12	07	
											Total
Hyp. ct.		16	15	20	24	23	18	19	26	25	186
Pct. ct.		52	48	64	77	74	58	61	84	81	66

TABLE 4

INTERCORRELATIONS AMONG SECOND-ORDER FACTORS ARRANGED TO SHOW APPROXIMATION TO HIERARCHICAL ORRER

Symbol	Gs	Gv	ER2	Gf	PSI	Gc	PRM	F	C2(−)
Gs									
Gv	37								
ER2	27	34							
Gf	39	34	17						
PSI	16	25	14	00					
Gc	10	36	−01	16	33				
PRM	22	−11	12	00	10	−01			
F	22	09	06	20	06	−12	04		
C2(−)	26	−02	19	−06	03	−12	09	−22	
Sums	1.99	1.62	1.28	1.20	1.07	0.69	0.45	0.33	0.15

$^{16}/_{31}$ths of the total variance. The results from these analyses are presented in a Table deposited with ADI.[2]

To achieve an approximation to a simple structure, Kaiser's (1958) Varimax procedure was employed and this was followed up with 13 graphic-plot visual rotations. The Varimax results have been deposited with ADI. The oblique simple structure solution is given in Table 3.

In Table 3, the values in the row labeled "Hyp. ct." are the actual numbers of loadings less than or equal to .10 or greater than or equal to −.10, and the values in the "Pct. ct." are the percentages of variables in the ±.10 hyperplanes for the factors.

The intercorrelations for the rotated factors are given in Table 4.

INTERPRETATION AND DISCUSSION OF RESULTS

The Correlations

Almost all of the intercorrelations among ability variables are positive. Exceptions are a −.10 correlation (not significant) between Intellectual Speed (Sp) and Carefulness (C) and a −.15 correlation (significant but low) between Mechanical Knowledge (Mk) and Writing Flexibility (Wf). Positive manifold for intercorrelations among abilities is, of course, a well-established finding.

Of particular interest in the present study are the correlations which factors like L, I, and CMR—that is, hypothesized Gf markers—have with other factors. L tends to be speed-free. It was obtained from responses to items arranged in sets of five in each of which there was a continuum from very easy to very difficult problems, and the score is the average level of difficulty of the problems solved, not merely the number correct: A person can obtain a very high score by working through only a few sets of five and getting them all correct, whereas he will get a low score if he works many problems but gives incorrect answers for some of them. Both I and CMR seemingly involve some speed variance, since they were given under timed conditions. Yet the pattern of correlations between L and other variables is very similar to the pattern of correlations of CFR and I with other variables. CFR and L, particularly, and I to a lesser degree, show relatively low correlations with the fluency factors (the average r is .18) and with measures of knowledge, per se (e.g., Mk, EMS, and V, where the average r is .22, as well as formal education: average $r = .21$). The correlations of L with I and CFR are both .53 and these latter two correlate .67. Thus, overall, the suggestion is that factors like I, CFR, and L definitely do involve intelligence, as putatively represented in other factors, but, relative to these other factors, are somewhat more free of formal accultural influences.

Many of the primaries said to indicate Gc—particularly V, EMS, and Mk—have positive correlations with both age and education, whereas some of the markers for Gf—notably CFR, Ma, and I—have negative correlations with age (−.29, −.16, and −.12 being the correlations for the three primaries mentioned) though they correlate positively with education (the values for the mentioned primaries being .19, .08, and .27). These findings thus again support the general hypothesis that the markers for Gf and Gc involve intelligence as this is generally understood.

---

[2] Material supplementary to this article has been deposited with the American Documentation Institute. Order Document No. 9029 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $1.25 for microfilm or $1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

The fluency factors, Fi and Fa, sometimes treated as indicators of "creativity" (cf. Getzels & Jackson, 1962; Guilford, 1962; Taylor, 1964), have rather low correlations with putative measures of intelligence (the average r for Fa and Fi being .18 with I, L, and CFR and .25 with R, V, Rs, and EMS). Thus, if one were to favor the currently popular viewpoint that creativity and intelligence are largely independent, he would seemingly find support for his position in these results. But Fa and Fi show rather high correlations (average .28) with simple Speed of Copying and Writing Flexibility, neither of which is regarded as a good measure of creativity (or intelligence). The suggestion is that a substantial portion of the reliable variance of Fa and Fi does not reflect creativity at all, but indicates motor speed, the rather simple crystallized skills of cursive writing, and the temperament factors associated with these measures (Fa and Fi correlate .22 and .21 respectively with UI-25, Exuberance). This interpretation is consistent with previous findings and probably needs to be considered within the context of a carefully argued theory of creativity, such as that which Burt (1962) has recently outlined.

The Factors

The "solidness" of the positive manifold among primary ability intercorrelations is made very evident in the first principal axis factor. This contains not only variables representing noegenetic processes in abstracting, reasoning, knowledge acquisition, etc., but also nonintellectual variables such as anxiety (U24), Radicalism (Q_1), Exuberance (U21), Self-Sentiment Development (U36), etc. Thus the factor represents a conglomerate of intellectual and nonintellectual functions.

As concerns the theory being developed in this study, the Varimax solution provides a confirmation of the general hypothesis, but a confirmation which is much clarified by the improved simple structure achieved by visual rotation. The first, very broad (poor simple structure) Varimax factor involves spatial visualization processes to a considerable extent, but it contains too much variance in nonvisual primaries (such as CMR, Rs, and L) to give really clear support for the hypothesis of a second-order general visualization dimension. With further shifts to the more nearly adequate simple structure provided by visual rotation,[3] however, the ambiguities inherent in the Varimax solution are largely erased, as is consistent with Thurstone's (1947) general theory of simple structure. The third factor in the visually-rotated solution provides quite definite support for the hypothesis of a general visualization dimension spanning Vz, S, Cf, Cs, and

---

[3] For readers not familiar with the procedures of visual rotation, it should be pointed that in the crucial meaning of the term "objectivity," this method of rotation is just as objective as analytic methods of rotation. All the rotator sees when he rotates are points and his objective is to shift factors so that a maximum number of points will have essentially zero projections on reference vectors drawn at right angles to corresponding factors. He is not aware of which variable corresponds to which point, just as a maximizing formula in analytic rotation is not thus "aware."

DFT, and dipping into measures of Gf in cases where these involve figural content. And what is illustrated with this example also applies to the other factors: The Varimax adumbrates a solution that is much more clearly indicated after visual rotation. Hence, although it is instructive to consider and compare the two rotated solutions, in the interest of efficient communication, only the visually rotated factors will be discussed in what follows.

The first factor in the oblique solution contains figural reasoning tasks (CFR), symbolic reasoning tasks (I, L, Sp) and semantic reasoning tasks (CMR, Rs, R). Intellectual Level (L), involving little or no speediness variance, is prominent in the pattern, but primaries involving intellectual speed, as best represented by Sp, are also present. The capacity to keep several fundaments in immediate awareness is represented in the loadings of Ma and P. Most important, the high-loaded variables are not such as reflect (in samples of adults) advanced schooling and other forms of intensive acculturation, whereas the variables which do most clearly reflect these influences are in the hyperplane of the factor. Hence this factor clearly corresponds to the hypothesized fluid intelligence function.

Factor II provides substantial support for the hypothesis specifying a crystallized function. Of the various primaries in this battery which may be said to involve intelligence to any considerable degree, practically every one which could be said to reflect intensive acculturation has a prominent loading in this dimension. The high loadings in Verbal Comprehension, Semantic Relations, General Reasoning, Formal Reasoning, etc., indicate that this is truly an intelligence dimension, involving noegenetic processes, but one which is very much nurtured by acculturation.

On first consideration it might seem anomalous that Factor II involves nonintellectual primaries such as UI-36 (Self-Sentiment), Q_1 (Independence) and UI-24 (Integration). But if we consider the kinds of influences which are said to result in Gc and those which produce the kind of self-sentiment and personality-integration indicated by the nonintellectual variables, it can be seen that the two overlap to a very great extent. It is difficult to specify which is the chicken and which the egg in this relationship, but the fact of the relationship itself is consistent with the general theory. In fact, this same kind of alignment of self-sentiment and Q_1 with a crystallized intelligence function was found in a previous study (Horn, 1961) aimed at demonstrating the mutually supportive interrelationship between ego functions, the self-sentiment and educated intellect.

McFarlane Smith (1965) has recently provided an excellent historical and theoretical account of the general visualization function, but the present results in Factor III would appear to provide the clearest verification yet presented of the thesis (dating back at least to El Koussy's, 1935, important results) that visualization abilities stand somewhat apart from the analytic reasoning and crystallized abilities of Gf and Gc. It can be seen here that Gv involves a kind of fluency in imagining the way objects may change as they move in space, maintaining orientation with respect to objects in space, keeping configurations in mind, finding the gestalt among disparate parts in a visual field, and maintaining a flexibility concerning other possible structurings of elements in space. It will be noted that virtually every task which may be said to involve spatial content has a loading in this factor.

But there is a noteworthy difference between the results predicted by Smith's analyses and those actually obtained here. Working mainly within the framework of research done in Britain—as summarized in Vernon's (1950) book, for example—Smith tended to see general visualization as an even broader factor than is here revealed. Indeed, Gf and Gv were not clearly distinguished in his treatment, whereas the present results show that although the two are cooperative in several important tasks (notably CFR), they are nevertheless linearly independent (though correlated) at a general level. It seems likely on the basis of these results, however, that particularly gifted mathematicians and scientists, such as Newton, would often be highly endowed with both Gf and Gv (not to mention Gc), as Smith, in effect, argues on the basis of his analyses of biographical and other data.

In Factor IV such variables as the speed of copying and perceptual speed suggest that the factor could represent a motivational, state-like function of effortfulness in the immediate testing situation. Introspectively it makes good sense that there is such variance. In previous research (Cattell, 1957) the personality factor UI-21, Exuberance, has been identified with such "competitiveness" in the immediate situation.

Factor VI is defined primarily by the fluency primaries, and these have relatively little variance in the speed component of Factor IV. However, the fluency primaries do have some variance in the Gc-reasoning factor. Also, in addition to sheer fluency, Factor VI brings in some variance on CMR. Possible interpretations of these findings have been suggested by Christensen and Guilford (1963), who indicated that a factor like VI can represent: (*a*) size of a reservoir of "ideas" or (*b*) speed in "calling up" the ideas. Of course, these two functions could be closely related. Regardless of which emphasis is taken in interpretation, the present findings indicate that the reservoir-glibness function of Factor VI is largely independent of Gc: Crystallized intelligence involves fluency, to be sure, as represented by the loadings of Fa and Fi, but the fluency primaries, particularly, and V and CMR to a lesser extent, measure a kind of reservoir flow in the use of concepts which is distinct from the thinking processes of Gc.

The remaining factors in Table 3 are of less concern in this study, although they are rather intriguing. Factor IX, for example, suggests that the carefulness factor, measured among intellectual tests, stands virtually alone at the second order in this rather broad sampling of both intellectual and nonintellectual primaries. The pattern of loadings is virtually identical to the pattern of correlations which Carefulness showed with other primaries. The suggestion is that Carefulness (in the space of these variables) is not a first-stratum factor, but one which belongs at the second order.

Factor IX is similar in some respects to a second-order superego factor found by Horn (1961) in a sample of ability and dynamic motivation variables. If carefulness is taken as an indication of superego development, the suggestion is that the superego is involved in performances in General Reasoning, Intellectual Level and Verbal Comprehension. It seems to oppose what may be slapdash speediness, as represented in Sp. Carefulness also relates to the academic kinds of values expressed in Q_1 (Experimenting). This also supports an interpretation in terms of superego development.

Factor V can be identified as the general dimension Horn (1963, 1964) has found among a wide array of questionnaire measures (including the 16PF and MMPI). This dimension is probably the empirical referent for what most investigators mean when they use the term "social desirability." But it is not the only questionnaire factor which might be given this name. At least two broad dimensions are required to account for the correlations between questionnaire items, tests, and factors. The one represented here is prominently loaded by the general anxiety factor (UI-24) and the threctia dimension (see Cattell & Scheier, 1961), but it also involves variance in extroversion and in measurements of moral-ethical values, as defined in the 16PF G factor.

Factor VII is defined essentially by three variables: (*a*) Premsia, UI-16, (*b*) Assertive Ego, and (*c*) Mk (−). It repeats the well-known association of Ego Assertiveness with Premsia. The association with low mechanical aptitude has likewise been very characteristic of Premsia. The values expressed in the factor are those of preferring what might be called the "good and delicate things in life"—that is, "good" books, "good" art, "good" music, "good" food, etc. The Premsic "sensitivity" is that of a person who prefers English to mathematics, "imaginative novels" to "realistic" accounts of military or political battles, brandy to beer, ballet to burlesques, etc. In Mk, on the other hand, the person shows himself (and the *him* is appropriate here) to know the difference between a U-bolt and a yoke, a timing gear and a crankshaft, a keyhole saw and a rip saw, etc.

Factor VIII is a specific loaded only by Early Risks. There would appear to be no need to discuss this factor at this time.

The Correlations between Second-Order Factors

It will be noted that a positive manifold exists among the first six factors and that positive correlations generally prevail for all of the second-order factors isolated in this study. Thus it would appear that the main general factors isolated in this study are not completely independent, and that a more general integrating principle operates among them. According to the theory outlined in this paper and expressed in more detail by Horn (1965b), the principle implied here is to be understood in terms of the interaction in development between various intellectual and nonintellectual traits of the person and in terms of the interaction of both educative and nutritive environmental conditions.

To note the positive manifold in Table 4 is not to say that a single factor will explain the covariance there represented. Hierarchical order is not by any means evident in the matrix. It is questionable whether such an order should be found, both on general theoretical grounds and, in the present instance, on technical grounds. As concerns the theoretical issues, consider the fact that the positive manifold includes the positive self-image dimension. It is reasonable that this should be correlated with Gf and Gc, surely, and it is not unreasonable that it should be correlated with Gv and Gs, but it does not follow that it should fall into a hierarchical order defining a single unitary function. Hierarchical order is a considerably more stringent requirement than is positive manifold. It implies one and, more importantly, *only* one influence or attribute in a set of observations; positive manifold indicates one influence, but not necessarily only one. While it is reasonable to argue that positive self-image stems from some of the same influences as Gc, Gf, etc., it is not so reasonable to argue that these various functions involve only one thing in common. The implication is that a multifactor theory is called for among these dimensions at the third-order.

However, the technical issue alluded to above introduces difficulties which make it awkward, at least, and perhaps impossible, to adequately explore possible multifactor hypotheses at the third order in the present study. Thurstone (1947) presented convincing statistical and logical arguments for the contention that simple structure factors are relatively invariant with univariate and multivariate bias in the selection of subjects (and the results in the present study appear to be consistent with his deductions). But by these same arguments he suggested that the correlations between factors will not be invariant and in fact will reflect sampling bias. Thus in the present instance, where it seems evident that there is some sampling bias, the table of intercorrelations between second-order dimensions is possibly quite ambiguous as concerns the third-order factors. For example, the correlation of Gv with Gf (and also with Gc and Gs) probably represents, in part at least, sampling bias in drawing male and female subsamples.

In general, then, although trends are indicated in Table 4, the differences between these correlations should probably be interpreted very cautiously.

Integration of Results

The findings, overall, thus indicate that a major proportion of the variance observed in intellectual performances can be understood in terms of six major functions:

1. Fluid intelligence, representing processes of perceiving relations, educing correlates, maintaining span of immediate awareness in reasoning, and abstracting in both speeded and unspeeded tasks of a relatively culture-fair kind but involving figural, symbolic, and semantic content.

2. Crystallized intelligence, representing similar processes of perceiving relations, educing correlates, etc., in speeded and unspeeded tasks involving various kinds of content, but tasks requiring considerable pretraining to acquire techniques representing the accumulated wisdom of a culture.

3. General visualization, representing processes of imagining the way objects may change as they move in space, maintaining orientation with respect to objects in space, keeping configurations in mind, finding the gestalt among disparate parts in a visual field, and maintaining a flexibility concerning other possible structurings of elements in space.

4. General Speediness, an attribute measured in simple writing and checking tasks requiring little in the way of complex relation-perceiving, but an attribute producing variance in the measure of most intellectual factors.

5. Facility in the use of concept labels, an ability to quickly bring words (i.e., concept labels) from long-term memory into immediate awareness, a facility which the evidence of this study suggests is largely independent of comprehension of the subtlety of the concepts themselves, as indicated in Gc.

6. Carefulness, a factor indicating unwillingness to make a mistake, as indicated by few wrong answers, but manifested in a wide variety of intellectual tests.

These general functions are positively interrelated and positively related to other variables indicating an optimistic view of self, personality integration, independence, and upwardly mobile social attitudes. Such positive manifold would thus seem to represent an interdependence between the attributes of the person, the influence of the environment upon the person and the person's affect on his surroundings.

REFERENCES

BURT, C. The structure of the mind; A review of the results of factor analysis. *British Journal of Psychology,* 1949, **19,** 176–199.

BURT, C. Critical notice: The psychology of creative ability. *British Journal of Educational Psychology,* 1962, **32,** 292–298.

CATTELL, R. B. *Personality and motivation structure and measurement.* New York: World Book, 1957, Pp. 492–499.

CATTELL, R. B. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology,* 1963, **54,** 1–22.

CATTELL, R. B., & SCHEIER, I. H. *The meaning and measurement of neuroticism and anxiety.* New York: Ronald Press, 1961.

CHRISTENSEN, P. R., & GUILFORD, J. P. An experimental study of verbal fluency factors. *British Journal of Statistical Psychology,* 1963, **16,** 1–26.

EL KOUSSY, A. A. H. The visual perception of space. *British Journal of Psychology Monographs Supplement,* 1935, No. 20.

FRENCH, J. W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs,* 1951.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Manual for Kit of Reference Tests for Cognitive Factors.* Princeton, N. J.: Educational Testing Service, 1963.

GETZELS, J. W., & JACKSON, P. W. *Creativity and intelligence.* New York: Wiley, 1962.

GUILFORD, J. P. Three faces of intellect. *American Psychologist,* 1959, **14,** 469–479.

GUILFORD, J. P. Creativity: Its measurement and development. In S. Parnes & H. Harding (Eds.) *A source book for creative thinking.* New York: Scribner, 1962.

GUILFORD, J. P. Intelligence: 1965 model. *American Psychologist,* 1966, **21,** 20–26.

GUILFORD, J. P., & MERRIFIELD, P. H. The structure of intellect model: Its uses and implications. *Reports from the Psychological Laboratory,* University of Southern California, 1960.

HORN, J. L. Structure among measures of superego, ego, and self-sentiment concepts. Unpublished master's thesis, University of Illinois, 1961.

HORN, J. L. Second-order factors in questionnaire data. *Educational Psychological Measurement,* 1963, **23,** 117–134.

HORN, J. L. Social desirability, evaluation and a hierarchical model for the organization of questionnaire responses. Paper read at the Annual Meeting of the Rocky Mountain Psychological Association, May 1964.

HORN, J. L. A rationale and technique for estimating the number of factors in factor analysis. *Psychometrika,* 1965, **30,** 179–185. (a)

HORN, J. L. Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities. Unpublished doctoral dissertation, University of Illinois, 1965. (b)

HORN, J. L., & CATTELL, R. B. Vehicles, ipsatization, and the multiple-method measurement of motivation. *Canadian Journal of Psychology,* 1965, **19**(4), 265–279.

HORN, J. L., & CATTELL, R. B. Age differences in fluid and crystallized intelligence. *Acta Psychologica,* 1966, in press. (a)

HORN, J. L., & CATTELL, R. B. Age differences in primary mental abilities. *Journal of Gerontology,* 1966, **21,** 210–220. (b)

HOTELLING, H. The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics,* 1940, **11,** 271–283.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika,* 1958, **21,** 187–200.

McNEMAR, Q. Lost: Our intelligence. Why? *American Psychologist,* 1964, **19,** 871–882.

SMITH, I. M. *Spatial ability: Its educational and social significance.* San Diego: Knapp, 1965.

TAYLOR, C. W. *Creativity: Progress and potential.* New York: McGraw-Hill, 1964.

THURSTONE, L. L. Primary Mental Abilities. *Psychometric Monographs,* 1938, No. 1.

THURSTONE, L. L. *Multiple-factor analysis.* Chicago: University of Chicago Press, 1947.

VERNON, P. E. *The structure of human abilities.* London: Methuen, 1950.

(Received September 16, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 5 271–278

SEMANTIC CONSTRAINT IN PAIRED-ASSOCIATE LEARNING[1]

WILLIAM D. ROHWER, JR., AND STEVE LYNCH
University of California, Berkeley

A factorial design was used to evaluate an explanation of the facilitory effect of verb strings on PA learning. The principal assumption was that verbs exert more constraint on subsequent words than do conjunctions and thereby limit the number of appropriate responses for a given stimulus. It was predicted that the discrepancy between the learning rates produced by conjunction and verb strings would disappear if the number of semantically appropriate responses for each was equated. This was accomplished through the use of a recognition task that was compared with the more typical PA recall task. In 3 replications of the experiment with a total of 250 5th- and 6th-grade Ss, the amount of facilitation produced by verb strings was as great in recognition as it was in recall tasks, leading to a rejection of the constraint explanation.

The rate at which paired associates are learned can be markedly increased by presenting the pairs in the context of a sentence (Jensen & Rohwer, 1963a, 1963b, 1965). More generally, the amount of facilitation produced by a meaningful, syntactically structured, verbal string depends upon the form class of the word intermediate between the two members of each pair (Rohwer, 1964). Specifically, Rohwer found that variations in learning rate were associated with the part of speech of the connective in each string; verbs produced the most rapid learning, followed by prepositions and conjunctions in that order. An explanation for the effect of the form-class variable was made in terms of semantic constraint. It was reasoned that verbs impose narrower limits on subsequent nouns in a string than do prepositions and that prepositions impose narrower limits than conjunctions. If so, the class of appropriate response nouns is smaller when the connective is a preposition than when it is a conjunction and still smaller when the connective is a verb. Finally, it was assumed that the probability of selecting a correct response for a given stimulus noun increases as the size of the class of appropriate response nouns decreases. From these assumptions the prediction followed that the presentation of paired nouns in the context of verbal strings would produce faster learning when the connective was a verb than when it was a preposition and that a preposition connective would produce faster learning than a conjunction.

The present experiment was designed to evaluate this constraint explanation by testing one of its implications for paired-associate learning. The materials were constructed to permit independent manipulation of the form class of connectives and the size of the class of response nouns. We expected that, if the size of the response classes was equated, the presentation of paired nouns connected by a conjunction would produce a rate of learning as fast as that produced by the presentation of the same nouns connected by a verb.

---

[1] This work was supported, in part, by a contract with the United States Office of Education (OE6-10-273) through Project Literacy. The report was prepared at the Institute of Human Learning, which is supported by grants from the National Science Foundation and the National Institutes of Health.

TABLE 1

STUDY-TRIAL MATERIALS: CONJUNCTION AND VERB STRINGS

List A	
Conjunction	Verb
The BEE and a DOG.	The BEE chases a DOG.
A LION and the HORSE.	A LION scares the HORSE.
The STICK or the COW.	The STICK hurts the COW.
A TRAIN and a CAT.	A TRAIN kills a CAT.
The ANT or a MOUSE.	The ANT bites a MOUSE.
A HUNTER and the SHEEP.	A HUNTER calls the SHEEP.
The FUR or the DEER.	The FUR covers the DEER.
The LIGHTNING or the GOAT.	The LIGHTNING surprises the GOAT.
A MAN or the FOX.	A MAN helps the FOX.
An AIRPLANE and a PIG.	An AIRPLANE wakes a PIG.
List B	
A SPIDER and a POLE.	A SPIDER climbs a POLE.
A WORKER or the STAIRS.	A WORKER paints the STAIRS.
A KNIFE and the SHIP.	A KNIFE carves the SHIP.
The STORM or a CHAIR.	The STORM breaks a CHAIR.
The GLUE or the LADDER.	The GLUE fixes the LADDER.
A BIRD or the FENCE.	A BIRD uses the FENCE.
The ICE and a SIGN.	The ICE cracks a SIGN.
The NAIL and a ROPE.	The NAIL holds a ROPE.
The PENCIL or a ROCK.	The PENCIL marks a ROCK.
The BROOM and a WALL.	The BROOM sweeps the WALL.

EXPERIMENT I

Method

Materials and design. All subjects (*S*s) were given the same task: to learn a list of 20 pairs of high-frequency nouns. Each pair was presented in the context of a meaningful, syntactically structured verbal string. Two sets of verbal strings were constructed: in one of these the form class of the connectives was conjunction; and, in the other, the form class was verb. With the exception of the connectives, the strings comprising the two sets were identical.

Each set of strings was subdivided into two mutually exclusive subsets of 10 strings each according to the rule that all response nouns within one subset should be semantically appropriate for every string in the subset and semantically inappropriate for all strings in the other subset. Thus, the 10 nouns within each subset were completely interchangeable. This rule, of course, only applies to the verb strings since conjunction connectives provide no semantic information about subsequent nouns. All of the response nouns in one subset (List A) were names of animate objects capable of autonomous locomotion. All of the response nouns in the other subset were names of objects incapable of autonomous locomotion (List B). The verb strings were constructed first and the conjunction strings were formed by substituting a coordinate conjunction (and, or) for each of the verbs. A complete listing of the study-trial materials may be found in Table 1.

The size of the class of response nouns was manipulated by varying the conditions of response mode. Half of the *S*s were asked to *recall* each response noun when presented with the appropriate stimulus noun and the other half were asked to *recognize* the appropriate response noun in a list of ten of the response nouns used in the experiment. The recall condition was essentially similar to the learning task in which the differences between the effects of conjunctions and verbs was initially detected. The constraint hypothesis asserts that under this condition the size of the response class, given verb strings, is substantially smaller than that for conjunction strings. In contrast, the recognition condition was designed to equate the sizes of the response classes for both the conjunction and verb strings. This was done by presenting *S*s in both the conjunction and verb groups with identical lists of 10 alternative response nouns for each stimulus noun presented. All 10 of the alternative nouns were semantically appropriate for the string in which the given stimulus noun was originally presented. For example, whenever a stimulus noun that had been paired with a response noun from the first subset was presented, the list of response alternatives consisted of the entire first subset. Thus, the verb strings provided no more semantic information as to which of the 10 response nouns was correct than did the conjunction strings. For the conjunction condition, these materials consisted of 20 2 × 2 transparencies bearing the 20 paired associates in the context of conjunction strings. Similarly, the materials for the verb condition consisted of 20 transparencies bearing the paired nouns in the context of verb strings.

In summary, a 2 × 2 factorial design was used in which the first factor was Form Class (Conjunction versus Verb) and the second was Response Mode (Recall versus Recognition). The prediction that followed from the constraint hypothesis was that verb strings would produce substantially more efficient learning than the conjunction strings only in the recall conditions, not in the recognition conditions.

Subjects. The total sample consisted of 60 sixth-grade children drawn from two local elementary schools, 40 children from one school and 20 from the other. Within each school, *S*s were assigned randomly to the four experimental groups; each group was comprised of 15 *S*s, 10 from one school and 5 from the other.

Procedure. The task was administered by means of a study-trial, test-trial method in which all *S*s were given a total of two study and two test trials. This method permitted the testing of 5 *S*s simultaneously. The study-trial materials were identical for both Response Mode conditions and varied only with respect to the Form Class factor. The conditions of response differed according to the character of the response booklets used: Each response sheet for the recall condition simply bore a horizontal line on which the *S* was asked to write his response; in contrast, each response sheet for the recognition condition bore a list of 10 alternative response nouns below which appeared the horizontal line on which the *S* was asked to write his chosen response. The stimulus words were projected successively on the screen and were identical for all four experimental conditions.

When *S*s entered the testing room, they were given response booklets for Trial 1 and the task was described to them. They were asked to memorize the 20 pairs of words in such a way that they could supply the second noun in each pair when shown the first. The procedure of presenting the pairs in the context of sentences or phrases was described and it was suggested that this might increase their ability to learn them. Both the study trial and the test trial were illustrated by two examples in which two sample strings were presented followed by the presentation of the stimulus noun from each string. The first study trial then began with the projection of each slide at a 4-second rate. The experimenter (*E*) read each string aloud as it appeared on the screen. After the last string had been shown, the *S*s were told to open their response booklets to the first page, look at the stimulus noun, and write their response in the space provided. The stimulus nouns were presented at a 15-second rate to permit the *S*s sufficient time to select their responses, write them in the booklet, and turn the page in preparation for the next stimulus. As each stimulus was projected, it was read aloud by *E*.

TABLE 2

Mean Numbers of Correct Responses as a Function of Form Class and Response Mode (Experiment I)

Response Mode	Form Class			
	Conjunction		Verb	
	M	*SD*	*M*	*SD*
Recall	13.60	6.87	21.47	7.86
Recognition	22.47	8.99	27.00	5.44

MSE (1,56) = 27.44.

At the end of the first test trial the response booklets were collected and those for Trial 2 were distributed. The procedure followed in Trial 1 was repeated to the completion of Trial 2 except that the strings and the stimulus nouns were presented in an order different from that used in Trial 1.

Results

Learning was measured in terms of the number of correct responses emitted during the two test trials. The mean numbers of correct responses for each of the four experimental conditions are presented in Table 2. These results were assessed by means of analysis of variance which revealed significant main effects associated with Form Class ($F = 10.51$, $df = 1/56$, $p < .01$), Response Mode ($F = 14.17$, $df = 1/56$, $p < .01$), and Trials ($F = 242.93$, $df = 1/56$, $p < .01$). The interaction crucial for the constraint hypothesis was not significant ($F < 1$), nor were any of the remaining tests in the analysis of variance. Verb strings produced substantially more efficient learning than did conjunction strings and the recognition mode produced more correct responses than did the recall mode, as expected. Even though the interaction of principal interest was not statistically significant, the difference between verb and conjunction strings appeared smaller under conditions of recognition than under conditions of recall as predicted by the constraint hypothesis. Given this, it was decided to replicate the experiment on *S*s expected to be more homogeneous with respect to learning rate than those sampled in Experiment I.

EXPERIMENT II

The second experiment to be reported consisted of an exact replication of Experiment I augmented by an additional condition of Response Mode. Furthermore, the experiment was performed on both a sixth-grade sample and a fifth-grade sample of children.

The additional experimental groups were added to provide a corollary test of the constraint hypothesis. Suppose it were shown that under the original recognition condition, the difference between verb and conjunction strings could be significantly reduced. Such a result would be consistent with the constraint hypothesis but it would be open to an alternative interpretation, namely, that some characteristic peculiar to the recognition procedure itself prevents the detection of a difference. For example, it has been shown repeatedly that performance measured in terms of recognition yields higher estimates of the amount learned or retained than when recall is required (Postman & Rau, 1957). Thus it might be that differences at high levels of performance are not comparable to differences at lower levels of performance. In order to clarify the interpretation of such results, a second recognition condition was added. In this case, however, the condition was designed to permit semantic constraint to operate and to produce again a difference between the conjunction and verb conditions.

The second recognition condition (Recognition 10/5) differed from the first (Recognition 10/10) only in the composition of the lists of response alternatives offered for each stimulus noun. In the Recognition 10/10 condition, all of the response alternatives were appropriate for both the conjunction strings and for the verb strings. In the 10/5 condition, however, the 10 alternative response nouns were chosen so that only five would be semantically appropriate for the verb condition while all 10 were appropriate for the conjunction condition. Thus, the prediction from the constraint hypothesis was that verb strings would produce more efficient learning in the 10/5 condition than the conjunction strings but not in the 10/10 condition.

Method

Materials and design. The task and the materials for the second experiment were the same as those for the previous experiment. The only exception was the composition of the response sheets for the 10/5 group. In the 10/10 condition, the 10 alternative responses for each stimulus noun were drawn from the same subset to insure that all would be semantically appropriate. In the 10/5 condition, the 10 response nouns consisted of five from one subset, the subset containing the correct response, and five from the other, inappropriate subset. For example, if the correct response noun were a member of the animate subset, five of the response alternatives were the names of animate objects and the other five were names of inanimate objects. The members of the two subsets were randomly ordered on each response sheet such that no clear grouping was apparent.

It was necessary to divide this experiment into two distinct 2 × 2 designs since the simultaneous analysis of all six experimental conditions involved the confounding of two factors: the mode of response (Recognition versus Recall) and, within the recognition conditions, the semantic appropriateness of the alternative responses (Conjunction versus Verb strings).

The first design was a replication of Experiment I with the addition of two classificatory factors used to increase experimental precision. In all, five factors were involved: Form Class (Conjunction strings versus Verb strings); Response Mode (Recall versus Recognition 10/10); Grades (Fifth versus Sixth); Reading Ability Levels (High versus Medium-High versus Medium-Low versus Low); Trials (1 versus 2). Performance scores from a standardized reading test, the Stanford Achievement Test, were available on all *S*s and this information was used to implement a levels design. The experimental prediction of principal concern was that the verb strings would produce more efficient learning than the conjunction strings in the recall but not in the recognition condition.

With one exception, the second design was the same as the first. The Response Mode factor involved a comparison of the Recognition 10/10 condition with the Recognition 10/5 condition. Thus the experimental groups in the Recognition 10/10 condition were involved in both analyses. For the second analysis the prediction consistent with the constraint hypothesis was that verb strings would produce faster learning than the conjunction strings in the Recognition 10/5 condition but not in the Recognition 10/10 condition.

Subjects. The total sample consisted of 48 sixth- and 48 fifth-grade children drawn from a local elementary school populated generally by homogeneously bright students who come from middle- and upper-middle class homes. Within each grade, *S*s were ordered with respect to reading achievement scores and then divided into quartiles of 12 *S*s each. Two *S*s selected at random from each quartile were randomly assigned to each of the six experimental groups. Each group was assigned eight *S*s from each of the two grades.

Procedure. The procedures followed in Experiment II were the same as those followed in Experiment I. The task was administered by a study-trial test-trial method and *S*s were tested in groups of eight. A 4-second presentation rate was used during the two study trials and a 15-second rate was used during the two test trials. The *E* again read the strings aloud to *S*s as they appeared on the screen during the study trials and read the stimulus words aloud as they appeared on the screen during the test trials.

TABLE 3

MEAN NUMBERS OF CORRECT RESPONSES AS A FUNCTION OF FORM CLASS, RESPONSE MODE, AND GRADE (EXPERIMENT II)

Grade	Response mode	Form class			
		Conjunction		Verb	
		M	*SD*	*M*	*SD*
5	Recall	14.63	3.58	20.63	6.87
	Recognition 10/10	19.00	4.14	30.88	2.53
	Recognition 10/5	21.00	7.65	30.63	4.57
6	Recall	13.75	4.92	28.00	4.38
	Recognition 10/10	24.12	5.89	29.38	4.03
	Recognition 10/5	19.75	5.78	32.25	3.06

MSE (1,31) = 12.41.

Results

As in Experiment I, the dependent variable in the present experiment was the number of correct responses emitted. The results are presented in Table 3.

Design 1. An analysis of variance was performed on the data produced by the Recall and Recognition 10/10 groups. The main effect for levels of reading ability was not significant ($F = 1.59$, $df = 3/31$, $p > .10$), nor were any of the interactions with levels, nor was the main effect for grades ($F = 4.13$, $df = 1/31$, $p > .05$). The main effect for trials was significant ($F = 258.82$, $df = 1/31$, $p < .0005$), but none of the interactions of other factors with trials was significant.

The factors of principal interest produced notably large effects. Response Mode differences favoring the Recognition 10/10 conditions were significant ($F = 28.02$, $df = 1/31$, $p < .0005$) and Form Class differences favoring the verb conditions were significant ($F = 56.26$, $df = 1/31$, $p < .0005$). But the interaction of Response Mode × Form Class, predicted by the constraint hypothesis, was not significant ($F < 1$). A higher-order interaction, however, required a reinterpretation of these effects: The interaction of Grades × Response Mode × Form Class was significant ($F = 8.91$, $df = 1/31$, $p < .01$). The form of the interaction was such that the prediction derived from the constraint hypothesis was confirmed by the performance of the sixth-grade *S*s and contradicted by the performance of the fifth-grade *S*s. That is to say, the predicted interaction of Response Mode × Form Class was significant in the sixth-grade samples ($F = 6.52$, $df = 1/31$, $p < .025$) but not in the fifth-grade samples ($F = 2.78$, $df = 1/31$, $p > .05$). Furthermore, the direction of this second-order interaction in the sixth-grade was opposite that in the fifth-grade. In the sixth-grade groups, the verb strings produced more efficient learning than the conjunction strings only for the Recall conditions and not for the Recognition 10/10 conditions. In contrast, the difference between conjunction and verb strings for the fifth-grade *S*s was larger under the Recognition 10/10 condition than under the Recall condition. In sum, the constraint hypothesis was firmly supported by the results in the higher grade and just as firmly denied by the results produced by the younger children.

Design 2. In the analysis of variance performed on the data produced by the Recognition 10/10 and Recognition 10/5 conditions, only two significant sources of variation emerged: Form Class ($F = 53.55$, $df = 1/30$, $p < .0005$); and Trials ($F = 173.21$, $df = 1/30$, $p < .0005$). Neither of the two interactions that might have provided support for the constraint hypothesis reached statistical significance. The interaction of Response Mode × Form Class was not significant ($F = 1.19$, $df = 1/30$, $p > .05$) nor was that of Grades × Response Mode × Form Class ($F = 3.64$, $df = 1/30$, $p > .05$), although the latter was relatively substantial. The simple effects of the interaction of Response Mode × Form Class was significant for Grade 6 ($F = 4.50$, $df = 1/30$, $p < .05$) but not for Grade 5 ($F < 1$). The latter result is, of course, consistent with the results reported in connection with Design 1, but nevertheless, provides only weak support (confined to the sixth-grade sample at that) for the constraint hypothesis.

No prediction relevant to this outcome had been made in advance and no convincing explanation could be developed post hoc. Before making a serious attempt to interpret the striking differences detected in Design 1 between fifth- and sixth-grade children, it was decided to replicate the entire experiment on other samples of fifth- and sixth-grade children drawn from a comparable population. To accomplish this purpose, Experiment III was performed.

EXPERIMENT III

Experiment III differed from Experiment II in only two ways. First, the classification factor of reading ability levels was not used since it had proved of little consequence in the previous experiment. Second, the 96 *S*s were drawn from a different school in another district because the population of fifth- and sixth-grade children in the district initially chosen had been virtually exhausted by the first two experiments. The school sampled in the present experiment was selected specifically because of its close comparability with the school sampled in Experiment II with respect to the tested intelligence and the social-class membership of its students. Apart from these two deviations, Experiment III was an exact replication of Experiment II.

Results

Once again the dependent variable was the numbers of correct responses emitted on Trials 1 and 2. The results are shown in Table 4.

TABLE 4

MEAN NUMBERS OF CORRECT RESPONSES AS A FUNCTION OF FORM CLASS, RESPONSE MODE, AND GRADE (EXPERIMENT III)

Grade	Response mode	Form class			
		Conjunction		Verb	
		M	*SD*	*M*	*SD*
5	Recall	16.88	6.83	25.25	8.29
	Recognition 10/10	19.63	7.05	26.63	6.37
	Recognition 10/5	20.50	6.97	28.75	6.07
6	Recall	18.63	8.48	30.13	4.22
	Recognition 10/10	16.25	7.21	31.25	4.30
	Recognition 10/5	17.50	6.61	26.75	6.52

MSE (1,56) = 22.57.

Design 1. A four-way analysis of variance was performed in which the principal sources of variance were: Grades (Fifth versus Sixth); Response Mode (Recall versus Recognition 10/10); Form Class (Conjunction versus Verb); and Trials (1 versus 2). The main effects for Grades and for Response Mode were not significant ($F < 1$ in both cases). The result for Grades is consistent with that found in the school used in Experiment II but the result for Response Mode is radically different. We have no explanation for the fact that the recognition conditions produced no more efficient learning than the recall in the present experiment.

The main effect for Form Class was once again significant, ($F = 41.20$, $df = 1/56$, $p < .0005$) and by itself accounted for more than 25% of the total variation observed. Clearly, the two different kinds of connectives produce reliably and substantially different rates of learning.

The only remaining effect of statistical significance was Trials ($F = 230.63$, $df = 1/56$, $p < .0005$); none of the interactions were significant sources of variance. The two interactions of principal interest in view of the results obtained in Experiment II, Response Mode × Form Class, and Grades × Response Mode × Form Class, each accounted for less than 1% of the total variation and produced F ratios smaller than 1. The conclusion from these results is clear: The constraint hypothesis finds no support whatever in the performance of *S*s in Experiment III.

Design 2. The results of the analysis of variance performed on the data contained in Design 2 were entirely consistent with those reported in connection with Experiment II. The only significant effects were those associated with: Form Class ($F = 37.57$, $df = 1/56$, $p < .0005$); and Trials ($F = 176.51$, $df = 1/56$, $p < .0005$). The main effects of Grades and Response Mode each produced an $F < 1$, as did the interaction of Grades × Response Mode × Form Class. The interaction of Response Mode × Form Class also accounted for an inconsequential portion of the total variation, 1% ($F = 1.18$, $df = 1/56$, $p > .05$). Again, Experiment III affords no evidential support for the constraint hypothesis.

DISCUSSION

The hypothesis of semantic constraint is not sufficient to explain the facilitation of learning produced by verb connectives. In only one of the times that the basic experimental design was repeated did the results lend support to the constraint explanation. Even when the size of the class of appropriate responses was equated for conjunction and verb conditions, verb strings produced substantially faster learning. Thus, the phenomenon remains to be explained.

Before dismissing the notion of semantic constraint, one of its hidden assumptions should be exposed and examined. For clarity, the learning process may be divided into an input phase, corresponding to the study trial in the present experiment, and an output phase, corresponding to the test trial. Semantic constraint, as presently formulated, should have its effect on performance during the output phase, that is, when *S* selects a response from among those available to him. It is at the time of recall that the verb might semantically limit the number of alternatives from which a response must be chosen. This may be an erroneous assumption. The principal locus of the effect of verb strings on learning may be the input phase rather than the output phase, just as variations in the accuracy of recall can be localized in the phase of original learning rather than in the conditions of recall (Olton, 1965). If this analysis were verified empirically, it would suggest that explanations of the facilitory effect of verb strings should refer to conditions which affect the acquisition process during the phase of initial input.

Two other features of the present results deserve special emphasis. The first concerns the power of verb strings to produce high levels of performance. In each of the experiments performed, the presentation of the paired associates in verb strings was sufficient to produce performance under conditions of recall that was as efficient as that produced by conjunction strings under conditions of recognition. Practically speaking, verb connectives can reduce the difficulty of a recall task to that of a normal recognition task. This may rightly be construed as a demonstration of the commonplace that the manner in which learning materials are presented to the learner, or the manner in which he operates on them, markedly affects learning efficiency.

Finally, the results of repeated replication of the same basic design in the present effort demonstrate another commonplace, namely, the too infrequently considered caution that a single experiment is not sufficient to establish a phenomenon nor to decide an issue. Had the present design been performed only once, on the sixth-grade sample in Experiment II, the conclusion implied about the constraint hypothesis would have been diametrically different from that which is now strongly indicated.

REFERENCES

JENSEN, A. R., & ROHWER, W. D., JR. The effect of verbal mediation on the learning and retention of paired-associates by retarded adults. *American Journal of Mental Deficiency*, 1963, **68**, 80–84. (a)

JENSEN, A. R., & ROHWER, W. D., JR. Verbal mediation in paired-associate and serial learning. *Journal of Verbal Learning and Verbal Behavior*, 1963, **1**, 346–352. (b)

JENSEN, A. R., & ROHWER, W. D., JR. Syntactical mediation of serial and paired-associate learning as a function of age. *Child Development*, 1965, **36**, 601–608.

OLTON, R. M. The effect of a mnemonic upon the retention of paired-associate verbal material. Unpublished doctoral dissertation, University of California, Berkeley, 1965.

POSTMAN, L., & RAU, L. Retention as a function of the method of measurement. *University of California Publications in Psychology*, 1957, **8**, 217–270.

ROHWER, W. D., JR. The verbal facilitation of paired-associate learning. Unpublished doctoral dissertation, University of California, Berkeley, 1964.

(Received October 11, 1965)

Journal of Educational Psychology
1966, Vol 57, No. 5, 279-286

SIZE-DISCRIMINATION LEARNING AS A FUNCTION OF MOTIVATION-HYGIENE ORIENTATION IN ADOLESCENTS[1]

H. CARL HAYWOOD AND THEODORE D. WACHS
George Peabody College

Adolescent Ss (N = 80) with below-average intelligence (IQ = 76) were selected for high-motivation or high-hygiene orientation using the Choice-Motivator Scale, and were given a visual size-discrimination problem to a dual criterion of 9 out of 10 correct trials on 2 successive days. The procedure was repeated for 80 similarly selected Ss of average intelligence (IQ = 105.5). At each intelligence level, motivation-oriented (MO) Ss learned more efficiently than did IQ-matched hygiene-oriented (HO) Ss on some of the scores. In the average-IQ groups, personality differences did not predict learning differences in original acquisition, but MO Ss relearned faster than did HO Ss. In the low-IQ groups, MO Ss required fewer total trials. More low-IQ HO Ss failed to learn the discrimination in the 200 trials given.

In the study of intellectual development and functioning, one of the most prominent sources of knowledge has been the study of discrimination learning in children, both normal and mentally retarded (Stevenson, 1963). Considering the amount of work done with discrimination learning, it is surprising that so little emphasis has been placed upon the relation of this learning process to personality variables. A survey of the literature from 1940 to the present revealed only nine studies investigating the relationship of personality variables to performance in discrimination-learning tasks. Corresponding to their prominence in other areas of learning, studies of manifest anxiety were most frequent (e.g., Farber & Spence, 1953; Hilgard, Jones, & Kaplan, 1951; Spence & Farber, 1954; Stevenson & Iscoe, 1956). Studies have also related need for achievement and discrimination learning (e.g., Estes, Miller, & Curtin, 1962; Miller & Estes, 1961) as well as psychopathic tendencies, as measured by the MMPI, and performance in a conditioned-discrimination problem (Warren & Grant, 1955).

Most relevant to the present study is the work of Terrell and his associates, who investigated the effects of varied motivational factors upon discrimination learning (Terrell, 1958; Terrell, Durkin, & Wiesley, 1959; Terrell & Kennedy, 1957). According to Terrell (1958), middle-class children learned more efficiently under intrinsic motivation than under extrinsic motivation; that is, these children learned for learning's sake or for the achievement of learning rather than for material incentives. Lower-class children learned more successfully with material incentives. Although Terrell related these differences to personality variables, his discussion was post hoc in that there were no independent measures of personality characteristics. Recent developments in personality research have indicated that Terrell may be quite correct in considering differences in children's performances on a discrimination-learning task as being due to the type of motivational system (intrinsic versus extrinsic) that characterizes the individual child.

---

[1] During the period of this study, both authors have received the support of a training grant (HD-43) from the National Institute of Child Health and Human Development, the second author as a research fellow in psychology and mental retardation. Parts of this research were accomplished under the auspices of the Institute on Mental Retardation and Intellectual Development at Peabody College, whose support derives from another NICHD grant, HD-973. The authors gratefully acknowledge the help of S. Joseph Weaver and Robert Cantrell for testing subjects, Bryan Shepp for his technical advice on discrimination problems, and Fred Austin for building the apparatus. Special thanks are due to the principal of Tennessee Preparatory School, Charles Neighbors, and his associates, Robert Birdwell and Joanne Cotton, for their tolerance of research and researchers.

The present research investigates the relevance of the motivation-hygiene concept formulated by Herzberg, Mausner, and Snyderman (1959) and extended by Herzberg and Hamlin (1961) to performance of children in a size-discrimination task. The motivation-hygiene concept asserts that an individual can make two types of adjustment toward his environment. The first, a hygiene-oriented adjustment, is extrinsic in nature: The hygiene-oriented (HO) individual is motivated by extrinsic environmental factors, for example, a need for security, a need for comfort, and the avoidance of stress and difficult tasks. The contrasting form of adjustment, the motivator-oriented adjustment, is intrinsic in nature: The motivator-oriented individual (MO) is motivated by intrinsic self-actualizing factors, for example, the desire for achievement, responsibility, and interest in work, a high sense of function pleasure (Herzberg & Hamlin, 1961), and a tendency to approach tension-inducing situations. The approach tendency of the MO individual and the avoidance tendency of the HO individual have been confirmed by Haywood and Dobbs (1964).

If basic learning tasks are indeed sensitive to individual differences in personality characteristics, it would be expected that the MO individual and the HO individual, because of their differing motivational systems, would perform differently on a discrimination-learning task. The HO individual, being less concerned with achievement and being less interested in the task than in the environment, would be less sensitive to cues in the task and more willing to give up trying since he would be getting little satisfaction from the task. The MO individual, being concerned with achievement in a task, and being interested in the task itself because of his sense of function pleasure, would be more sensitive to cues in the task and more willing to keep on trying to solve the problem. Hence, it would be expected that the discrimination learning of MO individuals would be more efficient than would the discrimination learning of HO individuals.

The present study has the secondary aim of exploring some of the environmental variables which might be expected to maximize the effects of individual differences in motivation-hygiene orientation. Working with schizophrenic inpatients, Sandvold (1962) has shown that motivator orientation can be increased by adding to the effort required to perform a task. For Sandvold's unimproved schizophrenics, the effort involved in the task had to be related to a goal in order for the full benefits of an MO orientation to be realized. In the present study, degree of effort and relevance of effort to task solution, as well as intelligence level, were systematically investigated to determine whether they interact with the motivation-hygiene orientation in influencing speed of learning on a discrimination problem. While it was expected that MO individuals would show faster learning than would HO individuals, no definite predictions were made for effort, relevance, or the possibilities of a differential effect of personality at different levels of intelligence.

METHOD

Subjects

Subjects (*S*s) were 160 adolescents at the Tennessee Preparatory School, a residential school for dependent children. All children in this school are given the Lorge-Thorndike intelligence tests as part of the school's entrance procedure. A high-intelligence group was constituted from all those children having an IQ of 95 or above on the Lorge-Thorndike; a low-intelligence group was constituted of those children having an IQ of 85 or below. All children at the school who met these criteria were given the Choice-Motivator Scale to be described below. Those scoring in either the top (motivator) or bottom (hygiene) quartiles of difference scores were used as *S*s, making a total of four groups of *S*s: high-intelligence MO children, high-intelligence HO children, low-intelligence MO children, and low-intelligence HO children. The groups were equated for intelligence level and chronological age, as shown in Table 1. The *S*s in each group were then randomly assigned to one of four experimental conditions: high effort-task relevant; high effort-task irrelevant; low effort-task relevant; low effort-task irrelevant. In each condition there were 10 MO *S*s and 10 HO *S*s.

TABLE 1

CHARACTERISTICS OF THE SUBJECT GROUPS

Groups	MO *S*s	HO *S*s
High Intelligence		
N	40	40
IQ	105.5	105.5
CA	14–8	14–1
Low Intelligence		
N	40	40
IQ	76.2	75.0
CA	14–11	14–10

Apparatus and Measuring Instruments

The motivation-hygiene orientation of each *S* was measured by means of the Choice-Motivator Scale (Hamlin & Nemo, 1962), which consists of 20 paired vocational choices. The *S* is instructed to indicate which of the pair he would rather be and then give the reason for his choice. The scoring is based upon systematic manifest-content analysis of the response given, and not the choice. The answers are classified into motivator responses (five categories), hygiene responses (six categories) or into ten other categories which are not used in the final scoring (Hamlin & Nemo, 1962). A difference score, obtained by subtracting the number of hygiene responses from the number of motivator responses, was used to classify *S*s.

The learning task was a size-discrimination problem with form and position irrelevant. The correct size, either large or small, was counterbalanced for each group. Stimuli were provided by two Inline Readout projectors (#10N83). Each projector contained four different stimulus forms in large and small sizes. These were combined to make 24 pairs of stimuli, the stimuli within each pair being different in both form and size. Because of the type of stepping switch used, two of the pairs were randomly chosen to be repeated at the end of the series, making a cycle of 26 presentations of stimulus pairs presented before the cycle repeated itself. Order of the stimulus pairs in the cycle was randomized with the limitation that the correct stimulus could not appear more than three times in succession on the same side. The correct choice was either the larger or smaller of the pair depending upon which was designated as correct by the experimenter. The projectors were mounted 9 inches apart on the back of a 12 × 28 inch wooden panel which was itself mounted on a wooden stand. Windows were cut in the face panel so the projector faces could be flush with the face panel. Two 12 × 20 inch boards and a 12 × 28 inch board were placed at the sides and top of the panel to form an enclosing shield. The whole apparatus was painted flat black. Below each window was an open-switch button which *S* pushed to indicate his choice. Between the windows and 2 inches above them was a small yellow light which was lit when *S* made a correct choice. The intertrial interval remained constant at 5 seconds, with the yellow light remaining on for 3 of those 5 seconds, if *S* made a correct choice. The complete operation was electrically programmed.

The task-effort manipulation consisted of having each *S* press upon a pedal with his feet while being tested on the discrimination problem. The pedal consisted of two boards, 9 inches square, which were painted flat black and hinged at one end. The top of the pedal was covered by a thin rubber sheet. The amount of pressure needed to depress the pedal was dependent upon the number of springs inserted between the two boards. Up to six springs could be inserted with each spring being calibrated at 5 pounds pressure. Depressing the pedal closed a switch located between the boards. This switch had connections leading back into the program board so that, depending upon the experimental condition *S* was in, depressing the pedal either started the discrimination apparatus (Task Relevant) or stopped a ringing bell located in the experimental chamber (Task Irrelevant).

Procedure

The discrimination apparatus and pedal were placed on a low table in a corner of the room facing a window which was covered by a black curtain. The program board was in an opposite corner behind a plywood screen which was painted flat black. The *S* was asked to sit before the discrimination apparatus and the following instructions were given:

> While you're here, we're going to play a couple of games. The first one is to see how strong you are. You press down on this pedal with both feet till you hear a click and hold it till I tell you to let up. (For each successive depression of the pedal, *E* said: Good, now try it again.)

Following these instructions, the experimenter ascertained the maximum pedal weight *S* could depress. For high-effort *S*s, 5 pounds was subtracted from this maximum down to a minimum pressure of 15 pounds. For those *S*s who could depress 30 pounds with little trouble the pedal was left at its maximum pressure. For low-effort *S*s, the pedal was kept at 5 pounds pressure. Upon completion of this procedure, the following instructions were read to all *S*s:

> This is a game in which I want you to choose between two shapes which will appear in these windows and press the button under the one you choose. You are to try to guess the right shape. If you press the button under the right shape, this yellow light will go on showing that you are right. If you press the button under the wrong shape, the light will not go on. Press the button only once, let off, and don't press it again until some more shapes appear in the windows. The aim of this game is to see how quickly you can learn to choose the shape that makes the light go on. Any questions?

For *S*s in the task-relevant condition, the following instructions were then read:

> In order to play this game, however, you must keep your feet pressed down on this pedal. As long as this pedal is down the machine will stay on and you can try to light the light. If you let up on this pedal, the machine will go off and will stay off till you press the pedal down again. OK, you press down on the pedal and I'll go next door and start the game.

For *S*s in the task-irrelevant condition, the following instructions were read:

> While you're playing the game, however, I want you to keep your feet pressed down on this pedal. If you let your feet up, a bell will start ringing to let you know that your feet are up and it will not stop till you press down again on the pedal. So keep your feet on the pedal while you try to light the light. You press down on the pedal now and I'll go next door and start the game.

Upon completion of these instructions, *E* left the chamber and went behind the plywood screen where he started the apparatus. As can be seen from this procedure, task relevance was defined in terms of whether the *S*'s effort was necessary or not necessary to continue the task.

Each *S* was given up to 100 massed trials per session to a criterion of 9 out of 10 correct, on the same problem, for two successive sessions. Each *S* was given only one session per day. Until criterion was achieved, the correct size remained constant. Immediately upon the *S*'s achieving criterion, the correct size was reversed and *S* was run to a criterion of 9 out of 10 correct in the reversal procedure. If *S* had not solved either the original problem or the reversal by the end of 200 trials, he was terminated. At the beginning of each new session, each *S* was asked if he remembered how to play the game. If he did not, the original instructions were repeated to him.

Analysis

Four measures were obtained for each *S*: (*a*) the number of trials required to reach the acquisition criterion of 9 out of 10 correct for 2 successive days on the same problem (total acquisition score; those *S*s who did not learn the problem were given a score of 200); (*b*) the number of trials required to reach the first criterion during acquisition (original acquisition score; only those *S*s who could learn were included in this analysis); (*c*) the number of trials required to reach the second criterion, on the same problem, when *S* was brought back the day after reaching the original acquisition criterion (task carryover score; only those *S*s who could learn were included in this analysis); (*d*) number of trials to reach criterion on reversal.

The data obtained from the two levels of intelligence were analyzed separately, so that each analysis was for a homogeneous level of intelligence. For each analysis, except reversal, a 2 × 2 × 2 factorial analysis of variance (Lindquist, 1953) was used. Reversal data were analyzed in a two-way analysis of covariance design (Winer, 1962).

RESULTS

Low-Intelligence Group

As expected, MO *S*s took significantly fewer trials to reach the total acquisition criterion than did HO *S*s ($F = 5.23$, $df = 1/72$, $p < .05$).

Analysis of the total acquisition scores showed no effect of task effort and task relevance upon rate of acquisition for *S*s in the low-intelligence group, nor did these variables interact with the personality dimensions. Analysis of the original acquisition scores and task-carryover scores showed no significant differences between MO *S*s and HO *S*s under any of the experimental conditions. Further inspection of the data indicated that these analyses were spurious because of the high proportion of HO *S*s who failed to reach either criterion, which necessitated their being dropped from the latter analysis. While 12.5% of the HO *S*s failed to learn the problem in 200 trials, only 2% of the MO subjects failed to learn over this period ($p <$ .07).

Because task effort and task relevance failed to show any effect in any of the acquisition analyses, these dimensions were collapsed and the reversal data were analyzed by a two-way analysis of covariance (Winer, 1962), with covariance adjustments for total acquisition scores. This analysis showed that, contrary to expectation, HO *S*s solved the reversal problem in significantly fewer trials than did MO *S*s ($F = 5.85$, $df = 1/66$, $p < .025$). The various scores for the low-intelligence group are summarized in Table 2.

TABLE 2

MEANS AND STANDARD DEVIATIONS ON ALL DEPENDENT VARIABLES FOR LOW-INTELLIGENCE GROUPS

Measures	MO *S*s	HO *S*s
Original acquisition: (Trials to criterion)	47.58 (34.30)	59.90 (42.80)
Task carryover score: (Trials to criterion)	19.26 (21.38)	23.97 (31.18)
Total acquistion: (Trials to criterion)	70.10 (47.23)	98.10 (61.55)
Reversal: (Trials to criterion)	51.97 (55.34)	32.50 (36.91)

Note.—Original acquisition plus Task carryover score will not equal Total acquisition since the latter included all *S*s, while only those *S*s who could learn the problem were included in the two former scores. Parenthetical entries are standard deviations.

High-Intelligence Group

Analysis of the total acquisition score for *S*s in the high-intelligence group showed no effects of the environmental manipulations and no significant differences between MO and HO *S*s. Since there were no differences in the proportion of MO and HO *S*s who failed to learn (5% and 2% respectively), it was concluded that the analysis of the original acquisition and task-carryover data would not be invalidated by the necessity of dropping *S*s in only one group, as happened in the low-IQ analysis. Analysis of the original acquisition score showed no effect of effort or task relevance and no differences between MO and HO *S*s; however, on the task-carryover score,

TABLE 3

MEANS AND STANDARD DEVIATIONS ON ALL DEPENDENT VARIABLES FOR HIGH-INTELLIGENCE GROUPS

Measures	MO *S*s	HO *S*s
Original acquisition: (Trials to criterion)	44.75 (33.53)	45.10 (36.31)
Task carryover score: (Trials to criterion)	11.30 (5.46)	19.65 (21.93)
Total acquisition score: (Trials to criterion)	63.32 (47.80)	64.82 (47.25)
Reversal: (Trials to criterion)	26.21 (34.51)	26.94 (36.84)

Note.—Original acquisition plus Task carryover score will not equal Total acquisition since the latter included all *S*s, while only those *S*s who could learn the problem were included in the two former scores. Parenthetical entries are standard deviations.

MO *S*s, when brought back to the same problem on the next day after reaching original criterion, took significantly fewer trials to reach criterion again than did the HO *S*s under the same conditions ($F = 5.26$, $df = 1/72$, $p < .025$). Again, because effort and relevance showed no effects, these dimensions were collapsed and the reversal data were analyzed in a two-way analysis of covariance (Winer, 1962) with covariance adjustments for total acquisition scores. No significant differences between MO and HO *S*s on reversal were found ($F = .008$, $df = 1/72$). The various scores for the high-intelligence group are summarized in Table 3.

DISCUSSION

Individual differences in motivational system were found to have a significant effect on some of the measures of discrimination learning, both in high-intelligence and low-intelligence groups. This was so in spite of the fact that the task, size discrimination, while adequate for the low-intelligence group, proved to be too easy for the high-intelligence group so that a ceiling effect was encountered in this group. The environmental variables (task effort and task relevance) did not interact with the motivation-hygiene dimension. The failure to find interactions with the environmental variables offers at least three possibilities of explanation: (*a*) differences in the populations sampled in the present study and in Sandvold's (1962) study; (*b*) the possibility that the effect of the environmental variables is not a real effect at all; and (*c*) the possibility that the present operations for effort and relevance may not be sufficiently strong to elicit measurable effects. It is still possible that under conditions of higher effort, these variables might show an interactive effect.

While there were no differences between MO and HO *S*s on the high-intelligence level under conditions of total acquisition, original acquisition, and reversal, the MO *S*s showed clear superiority under the task-carryover condition. Since the mean scores of the MO and HO groups on the measures revealing nonsignificant differences were in the predicted direction, it is possible that with a more difficult task the MO *S*s would also show superior acquisition. An alternative explanation is that with *S*s of average and above-average intelligence, the motivational system would not affect original learning; however, since MO *S*s are more interested in learning and more task-oriented (by definition), they would be expected to pay more attention to relevant cues and thus to be able to recall or resume a previously attempted task. This notion could be tested by use of a task more suitable to the intelligence level under study, in which both acquisition and recall would be scored. If this reasoning is correct, there should be no differences between MO and HO *S*s on original learning, but the MO *S*s should be superior on recall (as they were in the present study); however, if our lack of acquisition differences is due to a ceiling effect in the present task, it would be expected that MO *S*s would be superior in both acquisition and recall. The present data cannot answer these questions.

The superior learning of MO over HO *S*s in the low-IQ sample in total acquisition scores supports the utility of the motivation-hygiene dimension in studying basic learning processes, especially among mentally retarded *S*s. In the low-IQ group, the original acquisition and task-carryover scores in the present study were not considered valid because of the high proportion of HO *S*s who failed to learn at all. This result raises a question about the nature of the effects of the motivation-hygiene process. It may well be asked whether the results obtained were due to the superior performance of the MO *S*s, the very poor performance of the HO *S*s, or a combination of both. The fact that a high proportion of the HO *S*s were unable to learn the task at all gives minimal support to the second hypothesis.

The present study, as an exploratory survey of the effects of differing motivational systems upon learning performance, has yielded promising data. For both low-intelligence and high-intelligence groups, the significant findings were as expected from motivation-hygiene theory. The findings also suggested at least two new studies in this area, studies designed both to replicate the present findings and to explore the possible means by which the motivational systems studied have their effects upon behavior. The present data suggest that it will be profitable, in the study of the learning process, to pay more attention to those motivational variables that can be psychometrically indexed, thus making experimental capital of the fact of individual differences, rather than wishing they would go away.

REFERENCES

Estes, B. W., Miller, L. B., & Curtin, M. E. Monetary incentive and motivation in discrimination learning-sex differences. *Journal of Experimental Psychology*, 1962, **63**, 320.

Farber, I. E., & Spence, K. Complex learning and conditioning as a function of anxiety. *Journal of Experimental Psychology*, 1953, **45**, 120–125.

Hamlin, R. M., & Nemo, R. S. Self-actualization in choice scores of improved schizophrenics. *Journal of Clinical Psychology*, 1962, **18**, 51–54.

Haywood, H. C., & Dobbs, V. Motivation and anxiety in high school boys. *Journal of Personality*, 1964, **32**, 371–379.

Herzberg, F., & Hamlin, R. M. A motivation-hygiene concept of mental health. *Mental Hygiene*, 1961, **45**, 394–401.

Herzberg, F., Mausner, B., & Snyderman, B. *The motivation to work.* New York: Wiley, 1959.

Hilgard, E. R., Jones, V., & Kaplan, S. J. Conditioned discrimination as related to anxiety. *Journal of Experimental Psychology*, 1951, **42**, 94–100.

Lindquist, E. V. *Design and analysis of experiments in psychology and education.* Boston: Houghton-Mifflin, 1953.

Miller, L. B., & Estes, B. W. Monetary reward and motivation in discrimination learning. *Journal of Experimental Psychology*, 1961, **61**, 501–504.

Sandvold, K. D. The effect of effort and task orientation on the motivation orientation and verbal responsivity of chronic schizophrenic patients and normals. Unpublished doctoral dissertation, University of Illinois, 1962.

Spence, K., & Farber, I. E. The relation of anxiety to differential eyelid conditioning. *Journal of Experimental Psychology*, 1954, **47**, 127–134.

Stevenson, H. Discrimination learning. In N. R. Ellis (Ed.) *Handbook of mental deficiency*. New York: McGraw-Hill, 1963.

Stevenson, H., & Iscoe, I. Anxiety and discrimination learning. *American Journal of Psychology*, 1956, **69**, 113–114.

Terrell, G. The role of incentive in discrimination learning in children. *Child Development*, 1958, **29**, 231–236.

TERRELL, G., DURKIN, K., & WIESLEY, M. Social class and the nature of the incentive in discrimination learning. *Journal of Abnormal and Social Psychology*, 1959, **59**, 270–272.

TERRELL, G., & KENNEDY, W. A. Discrimination learning and transposition in children as a function of the nature of the reward. *Journal of Experimental Psychology*, 1957, **53**, 257–260.

WARREN, A. B., & GRANT, D. The relation of conditioned discrimination to the MMPI Pd. personality variable. *Journal of Experimental Psychology*, 1955, **49**, 23–27.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received October 12, 1965)

Journal of Educational Psychology
1966, Vol 57, No. 5, 287-296

GROUP STRUCTURE AND PRODUCTIVITY IN PROBLEM-SOLVING[1]

BRYCE B. HUDGINS AND LOUIS M. SMITH
Washington University

The 4 experiments were conducted in which elementary school children from middle and lower social classes solved arithmetic or social studies problems. General support was found for 3 hypotheses: (a) Group solutions to problems are not better than the independent solutions by the most able member of the group if he is perceived to be most able. (b) In arithmetic, when the able group member is not perceived as most able, the group will do better than the high-ability member. (c) There will be a shift in the group's perception of a low-status high-ability member if the group's scores are not better than the individual's. One of the most interesting and unanticipated findings occurred with the social studies task where the groups had lower scores than the high-ability member working alone.

The research reported in this paper was an extension and elaboration of some of the findings of an earlier study (Hudgins, 1960). In that investigation, which attempted to examine transfer of arithmetic problem-solving skills from group to individual situations, it was noted that subjects who had worked in groups did not solve problems more effectively in a subsequent individual setting although the groups had consistently produced more correct solutions than comparable individuals working alone. Informal observation suggested that within the groups subjects met with differential success in having their solutions to problems accepted, this success apparently depending upon the status ascribed to them by the other members of the group. Later analysis of the data suggested that the problem-solving scores of the groups were highly correlated with the performance of the most able control subject. Two of the central questions of the present research developed from an analysis of this earlier study; these were hypotheses relating the productivity of the group to its most able member, and to the accuracy with which his ability is perceived by other members of the group. An additional hypothesis was formulated as the experimenters sought implications of the first two for the functioning of small problem-solving groups.

THEORY AND HYPOTHESES

Theoretically, an understanding of these questions required a number of concepts; these are listed and defined briefly below:

High-ability subjects (H-Ab *Ss*): group members who are demonstrably superior on abilities relevant to the task.

Average-ability subjects (Av-Ab *Ss*): group members who are demonstrably less able than H-Ab *S* but who have some ability relevant to the task.

Task: school problems, arithmetic and social studies, graded in difficulty.

Status: perceived ability relevant to the task.

The reasoning which links our concepts proceeds as follows: assume that an ideal group of three persons is composed of one H-Ab *S* and two Av-Ab *Ss*. Also assume that the task of the group is a series of problems perfectly scaled for difficulty, so that the first item is of minimum difficulty, each subsequent item increases by one step in difficulty to the last, which is of maximum difficulty. In such a group, each item which can be passed by Av-Ab *S* can also be passed by H-Ab *S*, but the converse is not true. Suppose that the group is given a series of problems to solve, scaled as described above, and designated a, b, c, ··· *f* ··· *n*. Problem *f* represents the upper limit of Av-Ab *S*'s ability, and *n*, which is theoretically at least several steps advanced in difficulty, the upper limit of ability for H-Ab *S*. Then, any one of the three group members can correctly solve problems *a* through *f*, but only H-Ab *S* can function effectively beyond that point.

---

[1] The research reported herein was conducted pursuant to the terms of a contract with the United States Office of Education, Department of Health, Education, and Welfare (Hudgins & Smith, 1962).

Assume that environmental conditions demand concensus about a solution before it can be accepted. An unusual and a quite difficult role is imposed upon Av-Ab *S*. It is essentially a role which involves allowing H-Ab *S* to make his maximum contribution, but at the same time imposing upon him the obligation to demonstrate the adequacy of his solutions. Conversely, H-Ab *S* is in the difficult situation of having more of a contribution to make and with individuals who do not have ability to judge his contribution.

An intriguing question concerns the recognition of the most capable member of the group (H-Ab *S*). Drawing upon the earlier illustration, when Problem *g* and subsequent problems are reached, Av-Ab *S*s have no way of proposing workable solutions, yet H-Ab *S* has not reached the limits of his ability. When he suggests a solution, what conditions are likely to influence Av-Ab *S*s' acceptance of it? First, acceptance of his solutions should come more readily if H-Ab *S* had high status prior to the formation of the group; that is, if Av-Ab *S* has had earlier acquaintance with him as a bright student. Second, if H-Ab *S* is not initially perceived accurately, his solutions should meet with diminishing resistance, perhaps merely because he is the only member who can propose reasonable solutions to increasingly difficult problems. The probability that Av-Ab *S* will accept a solution to a problem in the *g* to *n* range would seem to be controlled by the *validation* procedure he employs in evaluating solutions. Since Av-Ab *S* is not able to evaluate H-Ab *S*'s solutions directly, that is, on their merits, he is likely to resort to a social criterion. It seems reasonable to anticipate that Av-Ab *S* is more likely to accept solutions offered by H-Ab *S* when the latter is perceived as one who typically knows the right answer to problems.

When H-Ab *S* is clearly perceived as the most able member of his group, one would expect the productivity of the group to approximate the productivity of H-Ab *S* closely, on the grounds that Av-Ab *S* recognizes H-Ab *S*'s superiority on the task dimension. For each of the four experiments reported in this paper, the first hypothesis reads: When the most able member of a small problem-solving group is also a high-status member of the group, the productivity of the group is a function of the productivity of the high-ability member. Operationally, this means that no statistically significant difference in the number of problems solved will occur between the situation where H-Ab *S* works alone and that in which he is a member of a small group.

If the high-ability member of a group is of low status, however, that is, if he is not perceived as the most able member of the group, the relationship between the high-ability member's productivity and that of the group described in Hypothesis 1 may not hold. If it does hold, there should be a change in the status of the high-ability member, such that he is now accorded high status by his group mates. Thus, Hypotheses 2 and 3 of the experiments bear a contingent relationship to each other.

The second hypothesis states: When the most able member of a small problem-solving group is also a low-status member of the group, the productivity of the group is not a function of the productivity of the high-ability member. Operationally this is interpreted to mean that the group will be significantly more productive than the high-ability but low-status member. In effect the group will demand that he indicate his solution processes and thereby render a checking and correcting influence.

The prediction in Hypothesis 3 is contingent upon the outcome of the second hypothesis. The third hypothesis states: If Hypothesis 2 is retained; that is, if the group is more productive than its most able member, the status of the high-ability member will not change. If Hypothesis 2 is rejected; that is, no significant difference results between group productivity and the productivity of the able member, his status will change from low to high. Operationally the hypothesis means that, in the first case, H-Ab *Ss* of low status will receive fewer nominations from their group mates as the most productive member of the group than high-status high-ability members will receive from their group mates. In the second case, that is, the case where Hypothesis 2 is rejected, the originally low-status high-ability members should be nominated as most productive about as frequently as high-status members are so nominated.

The theory and the hypotheses derived from it are built on the assumption that the tasks employed are perfectly scaled for difficulty. In Experiments 1A and 1B, arithmetic problems are used because, of all the curricular areas of the elementary classroom, arithmetic tasks probably come closest to meeting this assumption about scaling. But it seems also of interest and importance to know how significant this assumption is to the theory. If the theory accounts for group functioning when the assumption of scaling is fairly met, is it also useful when one can argue on a priori grounds that this assumption is violated, or at least met less well than in the case of arithmetic problems? Since curricular areas such as the social studies and language arts probably depart more from this assumption than do arithmetic and perhaps science, it seems to be important to know something about the generality of the theory to other areas. For that reason, Experiments 2A and 2B were conducted with social studies tasks.

Similarly, educational theory often has been extended uncritically to a variety of environmental contexts. Our interests in the broader urban community suggested that replications between lower socioeconomic schools and middle socioeconomic schools would be fruitful. Consequently we selected children from these two areas of the community.

METHOD

Sample

In order to test the hypotheses stated above, a total of 48 groups containing three members apiece were composed. Each group was given either an arithmetic or a social studies task, with the instruction to solve cooperatively as many problems as possible. The arithmetic task had 30 problems; the social studies task 35. Twenty-four of the groups were drawn from the schools in middle-class neighborhoods and 24 from schools in lower-class neighborhoods. Half the groups in each socioeconomic school sample worked on arithmetic, and half on social studies. The grade levels of the groups ranged from fifth to eighth grades.

All the groups contained one H-Ab *S* and two Av-Ab *S*s. Within each socioeconomic area, 12 of the groups contained H-Ab *S*s who were perceived as such by their peers. The H-Ab *S*s in the other 12 groups were not recognized by their group mates at the outset as superior in task-relevant ability.

Constituting the Groups

All three members of each group were drawn from the same classroom. To constitute the experimental groups, pupils in participating classes were administered the Arithmetic Reasoning section of the Stanford Achievement Test, Form J, or the Sequential Tests of Educational Progress, Social Studies, Part One, Form 3A. A variety of sociometric data was collected at the same time. For the experiment attention centered on nominations by classmates of the "three or four pupils in this class who are best in arithmetic (social studies)" and the "three or four pupils in this class who are poorest in arithmetic (social studies)."

For each class separately, means and standard deviations were computed for the test scores and the sociometric nominations on ability. First, in order to qualify as an H-Ab *S* a pupil had to score on the ability test one standard deviation or more above the mean for his class. Thus, if the class mean were 14.75 and the standard deviation were 3.25, a pupil must have a score of 18 or more to be identified as H-Ab *S*. The Av-Ab *S*s in each class were those who scored approximately at the mean. In this illustrative case, a pupil with a score of 14 or 15 would be so designated.

Second, whether a pupil identified by his ability test score as suitable would be included in the study depended upon meeting an additional criterion. The sociometric ratings of potential H-Ab *S*s were first inspected. If such a pupil received sufficient nominations as a good student to place him one standard deviation above the mean for his class, he was identified as a high-status H-Ab *S*. H-Ab *S*s who were rated below the mean were described as low status. Pupils who met the first criterion (test score) but not the second were dropped from the sample.

To be included in a group with a high-status H-Ab *S*, the Av-Ab *S*s must have nominated H-Ab *S* as one of the most able pupils in the group. For inclusion in a group with a low-status high-ability member, the Av-Ab *S*s must *not* have given him a positive nomination. In a few instances they nominated him negatively, although this criterion was not met rigorously.

Procedures

Each small group was removed from its regular classroom to a different location in the building where the experiment could be conducted without interruptions. Subjects were provided with experimental materials, and were told that they were to work together to try to solve the problems correctly. The only restriction imposed upon their method of operation was that a single answer to each problem was to be agreed upon by all members of the group before proceeding to the next problem. One or both of the authors was present at each group session. Each session was tape-recorded. In addition, the experimenter kept running notes on the behavior of the group members. The procedure followed was the same with all the groups, regardless of task or structure with regard to H-Ab *S*'s status.

The task was a set of 30 arithmetic problems (Stanford Achievement Test, Arithmetic Reasoning, Form K) reproduced by mimeograph to remove the impression of a standardized test,[2] or 35 social studies items, taken from Part One, Form 3B of the STEP Social Studies.

No specific time limit was set. Since groups obviously require longer to complete such a task than individuals do, it was intended to allow the groups as much productive time as they required. On the average, groups consumed about an hour. At the close of the experimental session, each subject was asked to "name the one person who helped the group the most to get the right answer. You may name yourself if you wish to do so." Responses were written privately, and were collected immediately by the experimenter. Subjects were assured that no one else would learn whom they had named.

RESULTS

The results are reported separately for each of the four replications of the basic design. The replications are labeled Experiments 1A, 1B, 2A, and 2B.

---

[2] Reproduction was done with permission from World Book Company, Yonkers, New York.

Experiment 1A (Arithmetic—High Socioeconomic Status)

Hypothesis 1 states that the productivity of a small problem-solving group in which the most able member is of high status will be no greater than the productivity of that individual member. Six groups were constituted to test this hypothesis. The pretest score of H-Ab *S* represents his individual productivity. The group's productivity is represented by the number of problems solved cooperatively by the three pupils in the group. In this case the mean pretest score for high-status H-Ab *Ss* was 19.83, and the average group score equal to 19.67. A test of the significance of the difference between correlated means resulted in a t of 1.54. For a two-tailed test with 5 df, this value does not approach significance. The hypothesis is retained.

The second hypothesis predicts that when the H-Ab *S* is of low status, the group will be more productive than the individual H-Ab *S*. As we have indicated, the basis for this prediction lies essentially in the fact that Av-Ab *Ss*, who are likely to be more dubious about the solutions offered by a member whom they do not recognize as able, will impose more rigorous checks on H-Ab *S* than they would a high-status member. In this case, if each of the members plays out his respective role, the result should be a higher score than H-Ab *S* would produce by himself. The pretest scores of the low-status H-Ab *Ss* in this experiment had a mean of 19.17. The mean group score was 20.67, just 1.50 problems greater on the average. The difference between the individual and group means yields a t of 1.97. A t of 2.02 is required for significance at the .05 level with 5 df for a one-tailed test. Although the obtained t is slightly lower than the required value, the authors' disposition is to retain the hypothesis; that is, to conclude for this replication that the groups were more productive than the individual H-Ab *Ss*. This interpretation is supported by the results of the next hypothesis, and by the later discussion bearing upon the relationship between the status of H-Ab *S* and the role he plays in the problem-solving group.

As explained earlier, all subjects were asked to indicate, immediately following the experimental session, which of the members contributed most to the group's success. These nominations were used as indications of the postexperimental status of H-Ab *S*. One prediction made by Hypothesis 3, and the one which seems appropriate here, states that if the groups are more productive than the low-status H-Ab *Ss*, a difference in status should remain between the high- and low-status H-Ab *Ss* after the experiment. The high-status H-Ab *Ss* received mean nominations of 1.83 (of a possible 2; self-nominations by H-Ab *S* were ignored) as most responsible for their group's success. The low-status H-Ab *Ss* received only .67 nominations on the average. This difference is highly significant $t = 5.80$, $df = 11$ for a one-tailed test; $p < .01$.)

Experiment 1B (Arithmetic—Low Socioeconomic Status)

For the lower socioeconomic status groups, the results of the first hypothesis are the same as those reported in Experiment 1A. The mean score for H-Ab *Ss* is 16; for the groups it is 17. This mean difference yields a nonsignificant t of 1.37. The groups do not differ in productivity from the able members, and the hypothesis is retained.

In the test of the second hypothesis, although the groups solve on the average 1.17 problems more than the low status H-Ab *Ss* alone (17.67 problems correct for the groups compared with 16.5 for the H-Ab *S*s), this difference is not significant. The *t* value is 1.80. The prediction for Hypothesis 3 thus is that the postexperimental status of low-status H-Ab *S*s will not be different from that of the high-status members. A mean difference of .17 exists between the two sets of scores. This difference favors the high-status members. They receive 1.67 nominations compared with 1.50 for the low-status H-Ab *S*s. However, the difference is not significant. The *t* of .42 suggests that the obtained mean difference between the two groups is not a reliable one.

Experiment 2A (Social Studies—High Socioeconomic Status)

In the higher socioeconomic status groups, the high-status H-Ab *S*s independently solved an average of 29.33 social studies problems. For the groups, the mean score was 27.17. This mean difference is not significant. It yields a *t* of 2.38 with 5 *df*. The hypothesis is retained.

In testing Hypothesis 2, when the same comparison is made between low-status H-Ab *S*s and their groups, the result is similar to that reported above. The H-Ab *S*s solve 29.17 problems on the average; their groups only 27.33. A *t* of 1.81, however, based upon the mean difference, is not statistically significant. The hypothesis is rejected since the groups are not more productive than H-Ab *S*s.

The prediction in Hypothesis 3 is that the postexperimental nominations to high- and to low-status H-Ab *S*s will not differ. The high-status members receive 1.5 nominations on the average, compared with 1.4 for the low-status members. This difference yields a nonsignificant *t* of .21, and the hypothesis is retained.

Experiment 2B (Social Studies—Low Socioeconomic Status)

The pretest scores of the high-status H-Ab *S*s have a mean of 22.17. The mean for their groups is 20.83. The difference, resulting in a *t* of .60, is not significant. The first hypothesis is retained.

When the high-ability members are of low status, their group's mean is 19.50. For the high-ability members, the mean is 21.17. The value of *t* in this comparison is 1.39, which is not significant. The second hypothesis, that the groups produce more correct solutions than the low-status high-ability members, is not retained.

The high-status members receive 1.33 nominations on the average from their group mates as most responsible for the success of the group. The low-status subjects receive only .50 nominations. The prediction in this case, Hypothesis 3, is that the mean nominations given to the high- and the low-status H-Ab *S*s should not be significantly different from each other. For this comparison, a *t* of 1.77 is obtained, which, for a two-tailed test with 11 *df*, is not a significant value. A *t* of 2.20 is required for significance at the .05 level. The third hypothesis is retained.

DISCUSSION

The results of the experiments provide reasonable support for the theory and hypotheses elucidated in the first part of this report. The interpretation is perhaps clearer and more comprehensible with respect to the arithmetic experiments than the social studies experiments. Two major aspects of the studies seem to require some discussion and explanation. One has to do with the change in status of low-status H-Ab *S*s. The other concerns the relative unproductivity of the social studies groups. Let us consider them in order.

A first concern regarding status might well lie with the sociometric measure of status, the perceived ability relative to the task. In an earlier paper (Hudgins, Smith, & Johnson, 1962) we found that the measures possessed reliability. A later investigation (Hudgins & Loftis, 1966) indicated that the "invisible child," the low-status high-ability pupil, does not interact with the teacher significantly less than visible children (high-status pupils), nor does he initiate less interaction with the teacher than does the visible child in the ongoing regular classroom.

However, when a small group of elementary school children are engaged in cooperative problem-solving, they must exchange information with each other about the probable correct response to each problem. Through these exchanges, the children presumably become aware to some extent of the contributions they and their group mates make to the total group operation. That is not to say that any of the members can explicitly identify the role or contributions of the others, but simply that, if asked, they can make a relative judgment about the roles played by the members of the group.

It would be expected that a high-status high-ability member would contribute a lion's share to the functioning of the group. He is already known to his group mates as the most able of the three. If he makes a central contribution to the group problem-solving, it should follow that he will be accorded appropriate recognition for his contribution.

But what of the low-status H-Ab *S*? Our reasoning suggests that he may be more closely examined by his group mates before they are willing to accept his solutions, and by this means the productivity of the group may be enhanced. However, since one or both of the Av-Ab *S*s must play a significant role to bring about this enhancement, it is not so clear that the role of the H-Ab *S* in this situation is the central one. Hypothetically when the group's productivity exceeds his, the postexperimental status of the low-status individual will not rival that of his high-status counterpart. If, however, the low-status member plays a more important role in furnishing solutions to problems, the score of the group should be similar to his own, and his postexperimental status should now be about the same as that of a high-status member. That is, there should be congruence between the roles played by the high-ability members and their group mates' acknowledgments of those roles.

To specify this relationship between contribution to the group and postexperimental status of the high-ability member as carefully as possible the protocols made during the group sessions were analyzed. Such protocols were available for 43 of the total 48 groups. For this analysis, the concept "contribution to the group" was defined operationally as "offering the solution to a problem which was accepted by the group as the correct answer." Two roles were identified: that of the H-Ab *S*, and that of the Av-Ab *S* who made the larger number of contributions, and who is subsequently referred to as the rival Av-Ab *S*.

In light of the previous discussion, the first comparison of interest is that between the high-status and the low-status H-Ab *S*s. Recall that in Experiment 1A, which was an examination of arithmetic problem-solving with higher

TABLE 1

COMPARISONS OF CONTRIBUTIONS TO GROUPS BY HIGH-ABILITY SUBJECTS AND RIVAL AVERAGE-ABILITY SUBJECTS

Experiment	H-Ab *S* (*M* No. of solutions)	Rival Av-Ab *S* (*M* no. of solutions)	SD_d	*t*
1A: Arithmetic—high socioeconomic schools				
High status	14.00	4.67	2.71	3.44**
Low status	10.83	10.50	3.01	.11
1B: Arithmetic—low socioeconomic schools				
High status[a]	17.25	7.00	3.02	3.39**
Low status[a]	12.25	5.00	2.18	3.33**
2A: Social studies—high socioeconomic schools				
High status	18.33	10.33	3.68	2.17*
Low status[b]	16.00	13.00	3.51	.85
2B: Social studies—low socioeconomic schools				
High status	13.33	12.83	1.84	.27
Low status	13.00	13.83	3.77	.22

[a] Based on analysis of 4 groups.
[b] Based on analysis of 5 groups.
* $p < .05$ (1-tailed tests).
** $p < .01$ (1-tailed tests).

socioeconomic status groups, the groups were no more productive than the high-status members, they exceeded the low-status members in productivity, and a significant postexperimental difference resulted between the two types of H-Ab *S*s, as predicted. When the contributions of the H-Ab *S*s are examined, it is revealed that the high-status subjects contributed an average of 14 solutions to their groups. Low-status subjects contributed 10.83, on the average, to their groups. This mean difference is not significant. With a standard error of 2.82, the *t* value is only 1.12. Thus, contribution to the group, defined as number of accepted solutions offered, does not seem to explain the residual status differences between high- and low-status H-Ab *S*s.

In the three other experiments, no significant difference occurred in the postexperimental status of the two H-Ab *S* groups. Analysis of the protocols with respect to contributions to group problem-solving by the high-ability members tends to confirm this lack of difference. None of the three comparisons is statistically significant.

For each experiment, the Av-Ab *S* who makes the more frequent contribution can be considered a rival to the H-Ab *S*. Presumably the closer his contributions rival those of H-Ab *S*, the fewer nominations H-Ab *S* will receive as most helpful to the group. The results of these comparisons are reported in Table 1. Notice that in the results of Experiment 1A, the rival Av-Ab *S*s contribute almost exactly as many solutions as the low-status H-Ab *S*s do. This is also true for both comparisons in Experiment 2B, which is an examination of Social Studies tasks with lower socioeconomic status groups.

The data do not allow an adequate explanation of the unproductivity of the social studies groups. Before entering into further discussion of the results of the social studies experiments, it seems advisable to make a general point about the performance of the groups working on social studies tasks. In most cases the groups solved correctly fewer problems than the H-Ab *S*s had solved when working alone. That was true for 16 of the 24 groups. Seven groups earned scores slightly higher than the able member, and for one comparison, there was no difference. The most obvious explanation of this unanticipated difference would seem to be associated with the tests themselves which were used as experimental materials. That is, if Form 3B of the STEP Social Studies Test (used by the groups) was more difficult than Form 3A which was used as a pretest, the differences would be explicable in terms of differential levels of difficulty. However, there seems to be little basis for such an explanation. Equal raw scores yield virtually identical converted scores for both forms of the test used in these experiments (Educational Testing Service, 1957). When the item difficulty levels are examined, 18 of the 35 items in Part One are more difficult on Form 3B, and 17 on Form A are more difficult (Educational Testing Service, 1957). Since all of the experimental groups tried all 35 items, it seems unlikely that lack of comparability between the two forms of the test could account for the obtained differences.

The findings do suggest that the task variable in problem solving requires analysis in terms of more genotypic constructs. For example, the differences between the two sets of experiments may be attributable to the more explicitly taught responses in arithmetic as opposed to social studies; or to special demands upon reading skills required in order to operate effectively with the social studies tasks.

The problem-solving model upon which this work is based suggests that the perceptions of subjects about the group tasks may be an important lead for future inquiry. That model not only assumes that tasks are perfectly scaled for difficulty; it presupposes that subjects' perceptions of the task as a scale are veridical. If the assumption of veridicality is untenable, as it very likely was in the social studies experiments, group productivity should be diminished accordingly. That is, in tasks where subjects do not perceive their limitations, they are less likely to be amenable to accepting solutions of others, because the faulty perception which leads a subject to believe he can handle the problem in the first place would be expected to militate against such recognition. Thus, H-Ab *S* and Av-Ab *S* might often arrive at conflicting solutions. Since Av-Ab *S* would have little if any opportunity to check the credence of his solution (as he has in a task like arithmetic) the group's final solution may be the result of a power struggle. If Av-Ab *S* is more ascendant than H-Ab *S*, or if the Av-Ab *S*s coalesce against H-Ab *S*, results very much like those obtained in these experiments would be predicted.

Explicitly, with respect to small group problem solving, this analysis points to further examination of the dynamics of groups engaged in problem-solving tasks whose substantive components, while requisite for appropriate solution, are still less clear than was true of the problems employed in these experiments.

REFERENCES

EDUCATONAL TESTING SERVICE, COOPERATIVE TEST DIVISION. *Sequential tests of educational progress.* (Technical report) Princeton, N. J.: Educational Testing Service, 1957.

EDUCATIONAL TESTING SERVICE, COOPERATIVE TEST DIVISION. *Sequential Tests of Educational Progress.* (*Teacher's guide*) Princeton, N. J.: Educational Testing Service, 1959.

HUDGINS, B. B. Effects of group experience on individual problem solving. *Journal of Educational Psychology,* 1960, **51,** 37–42.

HUDGINS, B. B., & LOFTIS, L. The invisible child in the arithmetic class: A study of teacher-pupil interaction. *Journal of Genetic Psychology,* 1966, **108,** 143–152.

HUDGINS, B. B., & SMITH, L. M. An analysis of laboratory and classroom group functioning. Final Report, Project No. 798, United States Office of Education, Cooperative Research Division, 1962.

HUDGINS, B. B., SMITH, L. M., & JOHNSON, T. J. The child's perception of his classmates. *Journal of Genetic Psychology,* 1962, **101,** 401–405.

THIBAUT, J., & KELLEY, H. H. *The social psychology of groups.* New York: Wiley, 1959.

(Received October 25, 1965)

Journal of Educational Psychology
1966, Vol 57, No. 5, 297-306

PREDICTIVE EFFICIENCY OF TWO MULTIVARIATE STATISTICAL TECHNIQUES IN COMPARISON WITH CLINICAL PREDICTIONS

C. E. ALEXAKOS[1]
University of Wisconsin

The college grade-point average of academically superior high school students was predicted by statistical methods (regression and discriminant analyses) on the basis of 19 independent variables, and by clinical counselors on the basis of all information collected through counseling interviews and testing during the 4 high school years. The obtained results indicated that the statistical were slightly superior to the clinical predictions; however, both types of prediction were not as efficient as one would wish them to be. Weighed combinations of the 3 methods increased the efficiency by 37% in the least favorable case and by 106% in the most favorable. It was concluded that combinations of clinical-statistical predictions could be more efficient and more useful in counseling college-oriented students than any single method.

Prediction of academic success beyond the high school level has been a major concern in educational and psychological research. Less than a dozen years ago Meehl (1954), comparing the efficiency of the statistical with that of the clinical predictions, pointed out that:

> for some reason the literature contains almost no carefully executed studies on the clinical-actuarial issue. Although a number of psychologists, psychiatrists, and sociologists have discussed the problem, empirical evidence concerning the relative efficacy of the two methods of prediction is largely wanting [p. 83].

Statisticians have developed various predictive techniques such as regression estimates, ratio estimates, logarithmic estimates, and even nonparametric methods of prediction. In educational and psychological research the most frequently used method of prediction has been the multiple regression analysis. By means of the "least squares" method the regression analysis takes into account the linear relationship among a number of concomitant variables in order to predict a criterion variable.

A multivariate technique which was developed almost simultaneously by Hotelling (1931), Fisher (1936), and Mahalanobis (1936), each working independently in different parts of the world, is the discriminant analysis. The discriminant analysis by maximizing the ratio:

$$\frac{\text{Between groups sum of squares}}{\text{Within groups sum of squares}}$$

seeks to maximize the distance between or among groups. The distance between pairs of groups may be colinear or not. This characteristic constitutes an advantage of the discriminant over the regression analysis, since the latter is restricted to linear relationships only. The method of discriminant analysis requires that measurements on the same variables are available for all members of the sample on which the discriminant functions are to be computed and all additional individuals that are to be classified into the different groups. By means of this method a given population or sample could be classified into two or more distinct and mutually exclusive groups on the basis of given measurements.

---

[1] Now at West Virginia University.

Rulon (Tiedeman, Rulon, & Bryan, 1951), while comparing the discriminant function with the multiple regression analysis, observed that:

> multiple correlation techniques are designed for and are essentially fitting for the study of relationships between measurements: that is, between graded classes. And the relationships so studied are generally holding among the members of a single class out of various interesting unordered classes.... The discriminant analysis does what the multiple correlation approach does not. It uses group membership as the criterion and makes all comparisons between groups and none within groups [p. 87].

Further, according to Tiedeman (Tiedeman et al, 1951):

> discriminant analysis might also be used as a test of whether the regression on a set of measurements is really linear or not. If classification according to the criterion is used to define the several groups, discriminant functions providing for maximum discrimination among these criterion classes in terms of the n measurements can be computed [p. 79].

It seems, therefore, that, while regression analysis cannot deal with discriminant analysis problems, discriminant analysis can be adapted in such a way as to deal with some regression problems, if the groups are defined in terms of the criterion variable.

The present study was designed to compare predictions concerning the college achievement of academically superior students made by regression equations, discriminant functions, and clinician counselors. Comparisons were made in terms of the obtained results and could possibly serve the purpose of accumulating empirical evidence concerning the relative predictive efficacy of each method. As was mentioned earlier, theoretical comparisons between regression and discriminant analysis have been made by Rulon and Tiedeman (Tiedeman et al, 1951), and between statistical and clinical predictions by Meehl (1954).

A word of caution is due here. The study is not intended to show that regression and discriminant analyses can be applied simultaneously and indiscriminately to the solution of all problems of prediction. As Rulon has pointed out, the two methods are designed to answer two *different types of problems*. In general, regression analysis produces better results and offers more information when the relationship between predictor variables and criterion is linear. Further, if the relationship between predictors and criterion is not linear, exponential curves of the independent variables can be fitted to improve and supplement the linear regression. However, as the degree of the predicting equation is raised, the computational labor increases disproportionately, particularly when the number of predictors is large. Thus, the number of terms needed for computing a second degree or parabolic equation is $n(n - 1)/2$ where n is the number of variables. When $n = 20$, the number of required terms $= 190$. To determine and evaluate the 190 terms requires very cumbersome, if not prohibitive, computations.

The discriminant analysis method, on the other hand, is designed to deal with classificatory problems and can be used only when the criterion variable is divided into classes, such as low, average, and high. This type of nominal classification, of course, is less refined than the interval scale used in regression problems. The results of the discriminant analysis, therefore, are likely to be coarser in the general case. Then, discriminant analysis can offer evidence concerning the linearity of the regression between predictors and criterion, and thus indicate whether a polynomial analysis is necessary or not. The present study exemplifies one specific case of the general problem, as discussed by Rulon and Tiedeman, (Tiedeman et al, 1951), and should not be confused with the general case.

METHOD

The study was conducted at the Research and Guidance Laboratory for Superior Students at the University of Wisconsin. Students who participate in the Laboratory program are selected by their teachers when they are in the ninth grade. The number of students each participating school sends to the Laboratory is proportional to high school size and ranges from 2 to 5 students per year. The selected students visit the Guidance Laboratory once a year while they are in high school and spend the day taking tests, completing information blanks and inventories, visiting university classes, and taking part in counseling interviews. For a more detailed description of the Laboratory operations the reader may refer to Rothney's (1961) article.

Two-year samples of Laboratory participants were used so that the regression equations and discriminant functions obtained by the first could be applied to the second. Male and female students were treated separately because previous analysis of data on Laboratory participants had shown significant sex differences in some independent variables. The first-year sample was composed of 163 students: 98 males and 65 females who graduated from high school in June 1962; the second of 125 students: 72 males and 53 females who graduated from high school in 1963.

The following data were available on all subjects:

1. Number of courses taken in each of five subject-fields during high school and grade-point average achieved in each field. (The ordinary 4-point scale was used to convert letter grades into quantitative points.)

2. Data relating to each student's family socioeconomic background were either obtained from the student's information blank or estimated by high school staff members. These were: (*a*) father's and mother's level of occupation, (*b*) father's and mother's amount of formal education, and (*c*) estimated family annual income.

3. Laboratory records contained each student's scores on the following tests: (*a*) Terman Concept Mastery, (*b*) Watson-Glaser Test of Critical Thinking, (*c*) Davis Reading Test—Level of Comprehension, and (*d*) Davis Reading Test—Speed of Comprehension.

Concerning the variables of father's and mother's level of education, each subject was assigned a numerical score on a scale from 1 to 8, the lowest score being assigned to parents who had not completed elementary school and the highest to parents who held advanced degrees. The variables of parental level of occupation were weighted on a scale from 1 to 6. High school faculties had classified each student's family income into one of three broad categories: below $5,000, between $5,000–10,000, and above $10,000.

Analysis of the data was performed by means of both multiple regression equations and discriminant functions for the year sample of 1962. The known grade-point average (CGPA) of the 1962 sample at the end of the second year in college served as the criterion variable in computing both the regression equations and the discriminant functions. The grade-point average of the 1963 sample was not known until after its prediction by the regression equations and the classification of the sample according to the discriminant functions. The criterion variable was internally related to the independent variables in the regression equations, but was externally used to establish the groups and the a priori proportions of students who should belong to each group in the discriminant analysis. Thus, the 1963 sample was used as a cross-validation sample in order to check the efficacy of the equations and functions obtained from the 1962 sample.

RESULTS

According to Multiple Regression

Multiple regression equations were computed on the basis of 19 independent variables to predict the grade-point average at the end of the second year in college. The 19 independent variables are presented in Table 4 along with the *t*-test values

TABLE 1

PERCENTAGE OF 1963 STUDENTS CLASSIFIED CORRECTLY INTO EACH GPA BAND

CGPA band	By regression		By discriminant analysis		By counselors	
	Male	Female	Male	Female	Male	Female
Below 2.00	70	25	60	57	30	25
2.00–2.99	90	53	66	41	38	35
3.00–3.49	30	55	45	85	55	50
3.50–4.00	00	33	8	33	31	25

computed on the beta weights and the critical ratio values of the discriminant function. Variables 10 (high school grade-point average in science courses) and 12 (mother's level of education) were significant predictors of the college grade-point average beyond the 5% level for male students. Variables 6 (high school grade-point average in English courses) and 17 (score on the Watson-Glaser Test of Critical Thinking) were significant predictors beyond the 2% level for females.

The multiple correlation coefficients were .74 and .69 for males and females, respectively. According to these coefficients it seemed that by using the obtained regression equations one would be able to predict the CGPA of similar samples of students at a rate significantly better than chance. The equations were actually applied to the 1963 sample and the obtained results are presented in Tables 1, 2, and 3. The predictions obtained by regression were point predictions, but for the purpose of comparisons with the classification made by the discriminant analysis tests and by the counselors the following criterion classes (or CGPA bands) were set up:

Band 1: CGPA less than 2.00;
Band 2: CGPA between 2.00–2.99;
Band 3: CGPA between 3.00–3.49;
Band 4: CGPA between 3.50–4.00.

According to Discriminant Analysis

As explained earlier, the discriminant analysis test can be used to classify a given population or sample to a number of distinct and mutually exclusive groups. The proportion of individuals falling into each group can be determined either empirically or a priori on the basis of theoretical knowledge. In the present example the proportions were established empirically on the basis of the CGPA of the 1962 sample which was known. The 1962 sample contained 10 male and 7 female students who had obtained CGPA below 2.00 (Band 1); 37 males and 27 females who had achieved CGPA between 2.00–2.99

TABLE 2

PERCENTAGE OF 1963 STUDENTS PREDICTED TO PERFORM AT A HIGHER LEVEL THAN THEY ACTUALLY DID

CGPA band	By regression		By discrimant analysis		By counselors	
	Male	Female	Male	Female	Male	Female
Below 2.00	30	75	40	43	70	75
2.00–2.99	3	47	31	35	65	59
3.00–3.49	0	25	10	5	30	10

TABLE 3

PERCENTAGE OF 1963 STUDENTS PREDICTED TO PERFORM AT A LOWER LEVEL THAN THEY ACTUALLY DID

CGPA band	By regression		By discriminant analysis		By counselors	
	Male	Female	Male	Female	Male	Female
2.00–2.99	7	0	3	24	3	0
3.00–3.49	70	20	45	10	15	35
3.50–4.00	100	67	92	67	62	75

(Band 2); 40 males and 22 females with CGPA between 3.00–3.49 (Band 3); and 11 males and 9 females with CGPA between 3.50–4.00 (Band 4).

According to the CGPA obtained by the 1962 sample, the proportions of individuals belonging to each band was computed as follows:

	Males	Females
Band 1:	10/98 = .10	7/65 = .11
Band 2:	37/98 = .38	27/65 = .41
Band 3:	40/98 = .41	22/65 = .34
Band 4:	11/98 = .11	9/65 = .14

These proportions served as the external criterion in the discriminant analysis test for computing the individual discriminant functions, on the basis of which the members of the 1963 samples were classified into one of the foregoing bands as additional data. That is, the classification of the 1963 samples was made according to the discriminant functions obtained from the 1962 samples and according to the foregoing empirical proportions.

The obtained overall discriminant functions were both statistically significant, the one for males at the .05 level and the one for females at the .01 level. The generalized distances between pairs of CGPA bands were significant between Bands 1 and 3, 1 and 4, and 2 and 3 for males, and between Band 1 and each of the other three for females. In view of these overall results, it appeared that the discriminant function of the female sample more nearly approached a linear model than did the male sample. The male discriminant function was rather curvilinear at the upper end because Group 4 was not significantly different than Group 2 while Group 3 was.

According to Counselors' Predictions

Counselors had been asked to predict each student's level of performance in college shortly before the student's graduation from high school. In doing so, they were guided by the following classification:

He or she will:
A. be eligible for major awards based on scholastic performance.
B. be eligible for minor awards based on scholastic performance.
C. complete training. Will not be eligible for awards nor on probation.
D. take longer than usual time to graduate though never on probation.
E. drop out of training for reasons other than scholastic difficulty.
F. be a marginal student. Will be on probation at some time.
G. drop because of low grades.

Although the foregoing descriptive categories do not seem to fit distinct CGPA classes, their meaning was thoroughly explained to predicting counselors in seminars and formal discussions conducted at the Guidance Laboratory. Thus, A meant that the student will be included in the Dean's list (GPA = 3.50 or bet-

TABLE 4

VARIABLES INVOLVED IN THE REGRESSION AND DISCRIMINANT ANALYSIS TESTS AND THEIR CRITICAL VALUES AS PREDICTORS

Variable	Regression		Discriminant Analysis	
	Males	Females	Males	Females
Number of English courses in high school	−.824	−.791	.361	.572
Number of mathematics courses in high school	−1.374	.930	.523	1.100
Number of foreign language courses in high school	.541	.280	.064	1.691
Number of social studies courses in high school	−.269	−.266	.408	.201
Number of science courses in high school	.547	1.108	.533	1.311
High school GPA in English	1.590	2.563*	.294	5.096**
High school GPA in mathematics	1.222	−1.059	.692	.100
High school GPA in foreign languages	.603	.709	.323	.416
High school GPA in social studies	1.095	−.716	1.082	.389
High school GPA in science	2.085*	1.029	.954	.859
Father's level of education	.469	−.044	.741	1.172
Mother's level of education	2.217*	−.655	3.052*	.139
Estimated family income	.891	−.372	.379	1.226
Father's level of occupation	−.710	−.042	.057	1.125
Mother's level of occupation	.327	−.664	.202	.218
Terman Concept Mastery score	1.234	.940	2.423	2.442
Watson-Glaser Test of Critical Thinking	.925	2.514*	.514	2.931*
Davis Reading Test: Level of Comprehension	−1.382	−1.066	.100	5.598**
Davis Reading Test: Speed of Comprehension	.056	.038	.114	1.458

Note.—In the regression tests $df = 78, 45$ for male and female students, respectively. In the discriminant analysis tests $df = 3, 76$ for males; 3, 43 for females.

* $p < .05$.

** $p < .01$.

ter), B that the student will retain a scholarship grant (GPA = 3.00 or higher), C and D = never on probation (GPA = 2.00 or higher), E, F, and G = potential dropouts or probationary students (GPA = below 2.00). The results of the three predictions are presented in Tables 1, 2, and 3.

Predicting counselors had available 4 years of data collected by the Laboratory during the students' annual visits. These data consisted of test scores, high school grades, information on students' family background, participation in activities and leadership roles, educational and vocational plans, and information collected through the counseling interviews with both students and their parents, as well as autobiographies and other samples of students' writing. That is, counselors had available all the measurements used in the statistical analyses plus various other types of information that were not included in the statistical predictions.

The statistical predictions were made on the basis of measurements on 19 variables, which are presented in Table 4. The 19 variables are presented in the order in which they were analyzed and not according to their significance as predictors. Also, in Table 4 the level of significance of each predictor variable is indicated. Student t tests were used to test the significance of the beta weights in the regression equations, and F tests in the discriminant analyses. The critical ratio values were printed in both cases on the output of the 1604 digital computer.

TABLE 5

HITS AND MISSES BY EACH PREDICTIVE METHOD AND CHI SQUARE TESTS BETWEEN PAIRS OF METHODS FOR THE 1963 SAMPLE OF MALE STUDENTS

CGPA band	Predictive method			χ_{12}^2	χ_{13}^2	χ_{23}^2
	Multiple regression	Discriminant analysis	Counselors			
1 Hits	7	6	3	.200	3.200	1.800
Misses	3	4	7	$p = .325$	$p = .078$	$p = .160$
N	10	10	10			
2 Hits	26	19	11	4.857	7.152	4.419
Misses	3	10	18	$p < .05$	$p < .01$	$p < .05$
N	29	29	29			
3 Hits	6	9	11	.960	2.556	.400
Misses	14	11	9	$p = .30$	$p = .10$	$p = .50$
N	20	20	20			
4 Hits	0	3	5	3.391	6.190	.722
Misses	13	10	8	$p < .10$	$p < .02$	$p = .50$
N	13	13	13			

Note.—$df = 1$ for all chi-square values.

Comparison Among the Three Types of Prediction

Comparisons of the predictive potency of each method with that of the other two were made in terms of the number of students who were classified correctly (hits) in each CGPA in contrast to those who were misclassified (misses) in light of each student's actual grade-point average at the end of the second year in college. The number of hits and misses by each method and the tests of significance between pairs of methods are presented in Tables 5 and 6 for male and female students, respectively.

From Table 5 it can be seen that the regression analysis was more effective than were the counselors ($\chi^2 = 3.20$; $p < .08$) in predicting the lowest band of CGPA. Regression predictions were also superior to those obtained by the other two methods in Band 2 ($\chi_{12}^2 = 4.857$, $p < .05$; and $\chi_{13}^2 = 7.152$, $p < .01$). The discriminant analysis was more effective than the counselors in Band 2 ($\chi_{23}^2 = 4.419$; $p < .05$).

In Band 3 the counselors predicted slightly better than the regression equation ($\chi^2 = 2.556$; $p < .10$), and in Band 4 the counselors predicted significantly better than the regression equation ($\chi^2 = 5.190$; $p < .02$). Also, the predictions by the discriminant analysis were slightly better than those of regression in Band 4 ($\chi^2 = 3.391$; $p < .10$).

Investigation of Table 6 reveals that the predictions made by the discriminant function were more effective than those made by the other two methods in Bands 1 and 3. Exact tests of significance (Ferguson, 1959, p. 173) produced the probabilities of .014 in Band 1. In Band 3 the χ^2 values were 3.199 ($.05 < p < .10$) between regression and discriminant analysis, and 5.729 ($p < .02$) between discriminant analysis and counselors. In Bands 2 and 4 the obtained results were so similar that no

TABLE 6

HITS AND MISSES BY EACH PREDICTIVE METHOD AND CHI SQUARE TESTS BETWEEN PAIRS OF METHODS FOR THE 1963 SAMPLE OF FEMALE STUDENTS

CGPA band		Predictive method			χ_{12}^2	χ_{13}^2	χ_{23}^2
		Multiple regression	Discriminant analysis	Counselors			
1	Hits	1	6	1			
	Misses	6	1	6	$p = .014$	$p = .50$	$p = .014$
	N	7	7	7			
2	Hits	9	7	6			
	Misses	7	9	10	.500	1.929	.121
	N	16	16	16	$p = .50$	$p = .20$	$p = .75$
3	Hits	11	16	9			
	Misses	8	3	10	3.199	.422	5.729
	N	19	19	19	$.10 > p > .05$	$p = .50$	$p < .02$
4	Hits	4	4	3			
	Misses	7	7	8	—	—	—
	N	11	11	11			

Note.—$df = 1$ for all chi-square values.

differences could exist between the three methods.

DISCUSSION

Statistical predictions depend mainly upon: (*a*) the overall similarity between the samples on which the equations or functions were computed and the samples to which they are to be applied for predictive (or cross-validation) purposes, (*b*) the degree to which the empirical data fit the linear regression model (goodness of fit), and (*c*) the amount of error involved in the measurement of both predictor and criterion variables. Clinical predictions depend, to a large extent, upon the global impression a counselee makes upon his counselor. Affected by the overall impression, the counselor tends to ignore some bits of information concerning his counselee and overweigh others. Meehl (1954, p. 92) presents evidence that "the clinicians systematically over-predicted grade average," and calls this tendency the "leniency error."

In the present study the statistical predictions appeared to be superior to those made by counselors, although they too were far from adequate, probably because the data did not satisfy all the aforementioned assumptions. The regression equation, for example, identified correctly more male students in the lower two bands of CGPA than in the upper two. This result seemed strange because, according to regression theory, the middle bands, or rather the students close to the mean CGPA, should have been predicted correctly more frequently. To obtain such a result, however, would assume identical distributions of the measurements on all 19 variables between the 1962 and 1963 samples.

Evidently this assumption was not satisfied by the data. Failure to predict the upper bands, while successfully predicting the lower ones, seems to imply that the measurements on the 19 variables for the 1963 male sample were, on the average, lower than those of the 1962 sample. For this reason, the regression equation would predict lower CGPA for the 1963 sample, and consequently miss the potential upper-band performers. Another cause contributing to this result could be the fact that the data of this example did not fit the linear regression model as was shown by the computed discriminant function.

In contrast to regression, counselors predicted more successfully the CGPA of the upper bands of male students, but at the same time overestimated the college performance of many students. This result seems to be in agreement with the studies reviewed by Meehl and supports his observation concerning the "leniency error."

For the sample of female students, no significant difference was observed between the predictions by multiple regression and clinician counselors. Significant differences, however, were found between the results of the discriminant analysis and the other two methods. This finding seems to imply that the college performance of females can be predicted more efficiently on the basis of group membership than by linear regression or clinical judgment. Particularly, in the lower band female students left college not because of poor academic performance but for other reasons. It seems from the results of the discriminant analysis that female students who dropped out of college had similar characteristics in both years and could be singled out and classified correctly on the basis of measurements on the 19 variables.

In summary, in spite of the better results obtained by statistical predictions, all three methods of prediction leave much to be desired in terms of reliability and accuracy. Statistical predictions could be improved by more precise and more reliable measurements as well as by the inclusion of more predictor variables in the matrix of linear equations. Counselors' predictions could be enhanced by a consistent effort to avoid "leniency errors."

TABLE 7

COMBINATION OF THE THREE METHODS TO OBTAIN A COMPOSITE PREDICTION

Predictive method	Average % efficiency		Relative % efficiency	
	Males	Females	Males	Females
Clinical	38.5	34.0	100	100
Multiple regression	48.0	41.5	125	122
Discriminant analysis	48.0	54.0	125	159
Composite	62.5	70.0	162	206

From the obtained results it appeared that combination of the three methods could produce more efficient predictions than any single method. For this purpose, each predictive method was weighed according to its average efficiency, and the results of Table 7 were obtained. It is evident from Table 7 that by a combination of the statistical-clinical methods a gain in efficiency ranging from 37 to 106% was obtained. This result seems to imply that a reconciliation between statistical and clinical methodology, instead of the present animosity, might serve better the cause of counseling students.

Each method has its own merits but none is flawless. Statistical analyses can combine the available measurements in a way that no clinician counselor can. Clinician counselors, on the other hand, can sense and take into account some individual characteristics of their clients that no statistical analysis can. Combinations of clinical-actuarial predictions, therefore, could be more reliable and more useful in the process of counseling.

In this study the regression analysis showed that for male students the variables of high school GPA in science and mother's level of education were the most significant predictors. The importance of the former variable can be explained by the observed trend among male students to major in the fields of science, mathematics, and engineering in college; the latter is more difficult to explain. It seems that mothers with a higher level of education exert some motivational influence on their sons to go to college and succeed. For female students the variables of high school GPA in English and the score on the Watson-Glaser Test of Critical Thinking proved the most significant predictors. This result can be explained in view of most female students majoring in the humanities, social studies, and kindred fields in college, which require high proficiency in English.

Similar results were obtained by the discriminant analysis test. However, the variable of high school GPA in science was not statistically significant in this test for male students, and the score on the TCM Test approached the .05 level of significance but was not large enough to be declared significant. The discriminant analysis on the female sample followed the same pattern as the regression analysis with the variables of GPA in English and the Davis Reading Test: Level of Comprehension being significant at the .01 level, the variable of the Watson-Glaser Test at the .05 level, and the TCM Test approaching the .05 level of significance. All significant discriminants for females were verbal variables. In view of these results, it seems logical to believe that proficiency in verbal and language areas is a strong predictor of college achievement for female students, but not so for male students.

REFERENCES

FERGUSON, G. A. *Statistical analysis in psychology and education.* New York: McGraw-Hill, 1959.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* 1936, **7,** 179–188.

HOTELLING, H. The generalization of "Student's" ratio. *Annals of Mathematical Statistics,* 1931, **2,** 360–378.

MAHALANOBIS, P. C. On the generalized distance in statistics. *Proceedings of the National Institute for Science and Industry (India),* 1936, **12,** 49–55.

MEEHL, P. E. *Clinical vs. statistical predictions.* Minneapolis: University of Minnesota Press, 1954.

ROTHNEY, J. W. M. A state-wide approach to the discovery and guidance of superior students. *School and Society,* Summer 1961, 271–274.

TIEDEMAN, D. V., RULON, P. J., & BRYAN, J. G. The multiple discriminant functions: A symposium. *Harvard Educational Review,* 1951, **21,** 71–95.

(Received December 13, 1965)

Journal of Educational Psychology
1966, Vol 57, No. 5, 307-310

COMPARISON OF THE EFFECTS OF POSITIVE AND NEGATIVE VICARIOUS REINFORCEMENT IN AN OPERANT LEARNING TASK

RICHARD KELLY
DePauw University

In this study 60 male and 60 female kindergarten children were tested in same-sex pairs with 1 S serving as a model and the 2nd S as an observer. ⅓ of the models performed a simple motor task under regular positive social reinforcement (praise), ⅓ under regular negative social reinforcement (criticism), and ⅓ received no reinforcement. After watching the models perform, the observers worked on the task without direct reinforcement. The results indicate that in both the direct and vicarious reinforcement groups the highest operant rates were maintained under the negative reinforcement and the lowest rates in the "no reinforcement" condition.

A number of investigations have been concerned with a comparison of the relative effects of reward and punishment on children's learning. These studies have been reviewed in articles by Kennedy and Willcutt (1964) and Marshall (1965).

There have been several recent studies dealing with observational or vicarious learning (Barnwell & Sechrest, 1965; Berger, 1961, 1962; Kanfer, 1965; Miller, Zavos, Vlandis & Rosenbaum, 1961; Sechrest, 1963). Two of these studies have included both reward and punishment conditions. Sechrest (1963) suggested that direct explicit positive or negative reinforcement to a performer represents an opposite implicit reinforcement to an observer who is simultaneously completing a similar task. Thus the subject (*S*) observing a positively-reinforced model is implicitly punished, while *S* observing a negatively-reinforced model is implicitly rewarded. On a subsequent task, therefore, the responses of the observer should not parallel those of the performer who has been directly reinforced. The findings indicated that observers' responses following reinforcement to a performer were opposite to those of the performers' responses. There were no significant differences between the groups that received explicit and implicit reinforcement but a reliable difference between the positively- and negatively-reinforced groups. In a second study (Barnwell & Sechrest, 1965) in which implicit reinforcement effects were ruled out by having the observer perform only after the reinforcement had been given to the model, first-grade *S*s responded to vicarious positive and negative reinforcements in a way which paralleled the performer's responses to direct reinforcement, while third-grade *S*s failed to respond differentially to the two types of reinforcements. Since the criterion was whether the observer chose to work on the same or a different task than the performer, no comparison of the relative effects of reward and punishment was possible.

The purpose of the present study was to compare the relative effects of positive and negative vicarious reinforcements in an operant learning task which precluded implicit reinforcement effects. It was hypothesized that the effects of vicarious reinforcement would parallel those of direct reinforcement. In an earlier operant learning study, Kelly and Stephens (1964) found that negative reinforcements given directly had a greater effect on response rate than positive reinforcements. A second hypothesis was, therefore, that the effects of vicarious punishment would be greater than reward.

METHOD

Subjects

The *S*s were 60 male and 60 female kindergarten children enrolled in four public elementary schools in Greencastle, Indiana. They ranged in age from 5 years, 5 months to 6 years, 9 months and were predominately from middle and upper-lower class families.

Apparatus

The apparatus, which was adopted from Stevenson and Fahel (1961), has been described in detail elsewhere (Kelly & Stephens, 1964). It consisted essentially of a box 31 × 30 × 4 inches which was placed on a table 28 inches high. At the front of the box was a bin 13¾ × 7½ × 2⅛ inches containing 750 marbles. To the right of the bin was a 9 × 13 inch Masonite tray with seven holes ¾ inch in diameter arranged in a geometric pattern. The *S*s stood in front of the apparatus directly across from the experimenter (*E*), who was seated behind the table. A runway carried the marbles that were dropped through the holes to a container in front of *E*. A screen 7 inches high across the middle of the box shielded *E*'s recording materials but left *E*'s head and shoulders visible to *S*.

Procedure

The children were tested in pairs with each *S* being randomly assigned a partner of the same sex. One member of the pair was designated as the performer (*P*) and the other as the observer (*O*). The *P* was instructed to stand in front of the apparatus while *O* stood to his left so that the responses of *P* could be easily observed. The *S*s were then given the following instructions:

> Now we are going to play a marble game. ______ (name of *P* inserted) is going to play first and then it will be ______'s (name of *O* inserted) turn. Let me show you how to play. Take a marble from here and drop it in a hole (*E* demonstrated by taking a marble from the bin and dropping it in one of the holes in the tray). You can use any of the holes you want but be sure to take one marble at a time and use only one hand. Now you try (*P* was given an opportunity to make a response). We won't use all the marbles that are here in the game. I will tell you when to stop and then it will be ______'s (name of *O* inserted) turn.

There were three learning conditions: (*a*) regular positive social reinforcement (praise), (*b*) regular negative social reinforcement (reproof), and (*c*) no reinforcement. The learning task, which required 6 minutes, was divided into 1-minute intervals. In the positive and negative reinforcement conditions, no reinforcements were given during the first 1-minute interval. Beginning with the second minute the first response after the first and thirtieth second of each minute was reinforced, giving a total of ten reinforcements during the learning task. In the positive reinforcement condition *E* made supportive or encouraging statements about *P*'s performance, while the reinforcements in the negative condition consisted of critical comments. Since the purpose of the study was a comparison of the effects of positive and negative reinforcement, the reinforcements were worded so that both reward and punishment would produce the same response tendency—increase in rate of response. This was accomplished by including in the reinforcing comments the implication that speed of response was important. For example, the positive reinforcements included statements such as "Good—that's fast," while the negative reinforcements included such comments as "That's slow—not too good" or "That's not very fast." A reference to speed was included in the first reinforcement in each 1-minute interval. In the control group *P* performed the task for 6 minutes without receiving any reinforcements.

When *P* had completed the task, *S*s were instructed to change places and *O* was allowed to play the game for 6 minutes. During the time *O* performed the task no reinforcements were given by E.

At the end of the experimental session both *S*s were commended for their performance.

RESULTS

The number of responses (marbles dropped through the holes by *S*) per 1-minute interval was used as the criterion measure. The data for the groups receiving the direct and vicarious reinforcement were analyzed in two separate 2 × 3 × 6 repeated measures analyses of variance (Winer, 1962).

An analysis of variance of the data for the groups receiving direct reinforcement showed a highly significant Reinforcement main effect ($F = 11.42$, $df = 2/54$, $p < .01$). The curves shown in Figure 1 indicate that the group receiving negative reinforcement showed the highest overall response rate after the first minute, while the control group responded at the lowest rate. The Sex and Trials main effects were not significant, nor were there any significant interactions ($p > .05$).

The analysis of variance of the data for the vicarious reinforcement groups indicated that the Reinforcement main effect was highly significant ($F = 7.13$, $df = 2/54$, $p < .01$). Figure 2 indicates that, as was the case in the direct reinforcement condition, the highest response rate occurred in the group which received negative reinforcement, followed by the positive reinforcement group, and then the control group. There were no other significant main effects or interactions ($p > .05$).

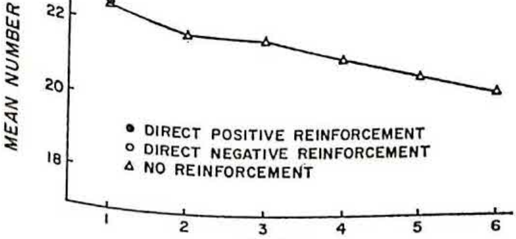


FIG. 1. Mean number of responses of groups receiving three types of direct reinforcement.

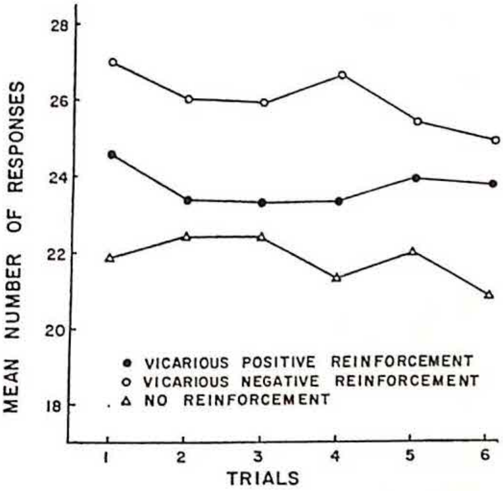


FIG. 2. Mean number of responses of groups receiving three types of vicarious reinforcement.

DISCUSSION

The performance of the groups receiving direct reinforcement in the present study provides a replication of the earlier Kelly and Stephens (1964) findings to the extent that there was a greater response to negative reinforcement than to positive reinforcement.

The responses of the vicarious reinforcement groups support the experimental hypotheses. The responses of the *O*s in both the positive and negative reinforcement conditions paralleled those of the *P*s in that both groups showed a higher response rate than did the control group. Also, as predicted, the response of the *O*s in the negative reinforcement condition was greater than that of the *O*s in the positive reinforcement condition. The parallel response of the *O*s is consistent with the Barnwell and Sechrest (1965) findings with first-grade children and lends some further support to the hypothesis that implicit reinforcement effects occur only when *O* is performing a task concurrently with *P*.

One possible interpretation of the present results is in terms of an expectancy hypothesis. The *O*s, by viewing the model perform the experimental task, obtained information as to the correct or desired response and also observed the type of reinforcements (i.e., positive or negative). It may be that as a result of this the *O*s developed a set of expectancies for either reward or punishment in the situation, and the effect of these expectancies was similar to that of the direct reinforcement. That is, expectancy for punishment had a greater effect on performance than expectancy for reward.

An alternative interpretation is that the punishment administered to the model served to increase the drive level of the model and also raised the drive level of the *O* who identified with the model. The heightened drive level would account for the more rapid response rate in the group receiving direct negative reinforcement. Furthermore, since *O* performed immediately after *P*, it would appear reasonable to assume that an increase in drive level as the result of identification with the model would carry over into the test situation. The higher drive in the vicarious negative reinforcement group would then explain the greater response rate of this group.

Several studies have suggested that punishment administered to the child can have detrimental side effects (Bandura & Walters, 1959; Sears, Maccoby, & Levin, 1957). Since the present findings indicate the effects of vicarious punishment parallel those of punishment administered directly, the question is raised as to whether detrimental side effects are also produced in the child who observes another being punished.

REFERENCES

BANDURA, A., & WALTERS, R. *Adolescent aggression*. New York: Ronald Press, 1959.

BARNWELL, A., & SECHREST, L. Vicarious reinforcement in children at two age levels. *Journal of Educational Psychology,* 1965, **56,** 100–106.

BERGER, S. M. Incidental learning through vicarious reinforcement. *Psychological Reports,* 1961, **9,** 477–491.

BERGER, S. M. Conditioning through vicarious instigation. *Psychological Review,* 1962, **59,** 450–466.

KANFER, F. H. Vicarious human reinforcements: A glimpse into the black box. In L. Krasner & L. P. Ullman (Eds.), *Research in behavior modification*. New York: Holt, Rinehart & Winston, 1965, Pp. 244–268.

KELLY, R., & STEPHENS, M. W. Comparison of different patterns of social reinforcement in children's operant learning. *Journal of Comparative and Physiological Psychology,* 1964, **57,** 294–296.

KENNEDY, W. A., & WILLCUTT, H. C. Praise and blame as incentives. *Psychological Bulletin,* 1964, **62,** 323–332.

MARSHALL, H. H. The effect of punishment on children: A review of the literature and a suggested hypothesis. *Journal of Genetic Psychology,* 1965, **106,** 23–33.

MILLER, G. R., ZAVOS, H., VLANDIS, J. W., & ROSENBAUM, M. E. The effects of differential reward on speech patterns. *Speech Monographs,* 1961, **28,** 9–15.

SEARS, R. R., MACCOBY, E. E., & LEVIN, H. *Patterns of child rearing*. Evanston: Row, Peterson, 1957.

SECHREST, L. Implicit reinforcement of responses. *Journal of Educational Psychology,* 1963, **54,** 197–201.

STEVENSON, H. W., & FAHEL, L. S. The effect of social reinforcement on the performance of institutionalized and noninstitutionalized normal and feeble minded children. *Journal of Personality,* 1961, **29,** 136–147.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received December 15, 1965)

Iowa Tests of Educational Development

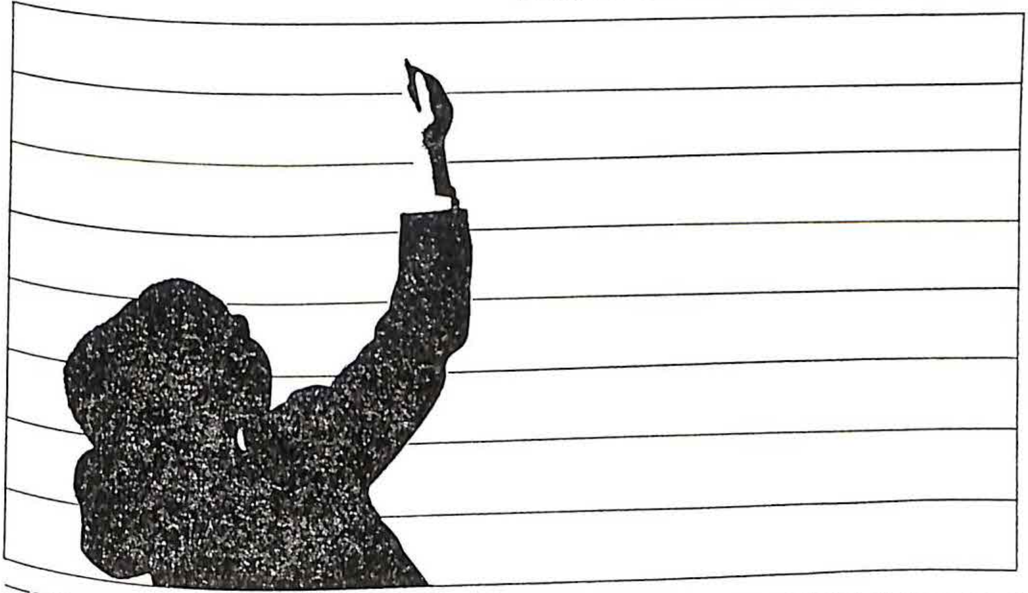
ITED gives you more information than it did last year.

The tests still number nine, but there are new ITED features that give you more data about the achievement of your high school students than you ever had before. And, interpretation can be more extensive.

Take a look at the newly designed Pupil Profile Report Form. Now student scores are plotted for you in the form of Confidence Bands. These bands tell you more about the relationships of scores than the old chart-yourself form could. In addition to the standard scores and national percentiles, you'll find corresponding local percentiles.

Another new feature is the Reading Average—the eleventh score to be reported by ITED. This average reading score is included on all reports to give you a single index of each student's reading achievement.

Other new features include an Item Analysis and a Research Plan. Both reports are highly technical to provide data for extensive research.



Testing this year? Return this coupon for a new brochure about ITED.

SRA

Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A Subsidiary of IBM

Name__________

Position__________

School__________

Address__________

City__________ State__________ Zip__________

from
RANDOM HOUSE

RANDOM HOUSE STUDIES IN EDUCATION AND PSYCHOLOGY

AUTOMATION, MANPOWER, AND EDUCATION
by JERRY M. ROSENBERG, *Teachers College, Columbia University*
September, 1966; SED5; 192 pages; $1.95

GUIDING THE GIFTED CHILD
by GEORGE I. THOMAS, *State Department of Education, Albany, New York, and* JOSEPH CRESCIMBENI, *Jacksonville University*
August, 1966; SED4; 192 pages; $1.95

ON THE IMPORTANCE OF INFANCY
by LAWRENCE K. FRANK PP32; 192 pages; $1.95

LEARNING Reinforcement Theory
by FRED S. KELLER, *Columbia University* PP13; 47 pages; $.95

THEORIES OF ADOLESCENCE
by ROLF E. MUUSS, Goucher College PP22; 190 pages; $1.95

For further information, write to,
RANDOM HOUSE, INC.
The College Department 501 Madison Avenue New York 10022

BORZOI College Textbooks

EDUCATIONAL PSYCHOLOGY
An Application of Social and Behavioral Theory
by LOUIS M. SMITH *and* BRYCE B. HUDGINS *both of Washington University at St. Louis*

Treating the principles and substantive findings of educational psychology in a manner that is at once scholarly, practical, and thoroughly sound, Professors Smith and Hudgins present an integrated view of social and learning behavior in the classroom.

"An excellent text! I am particularly appreciative of the incorporation of material from social and behavioral theory."
Sidney Yudin, *New York University*

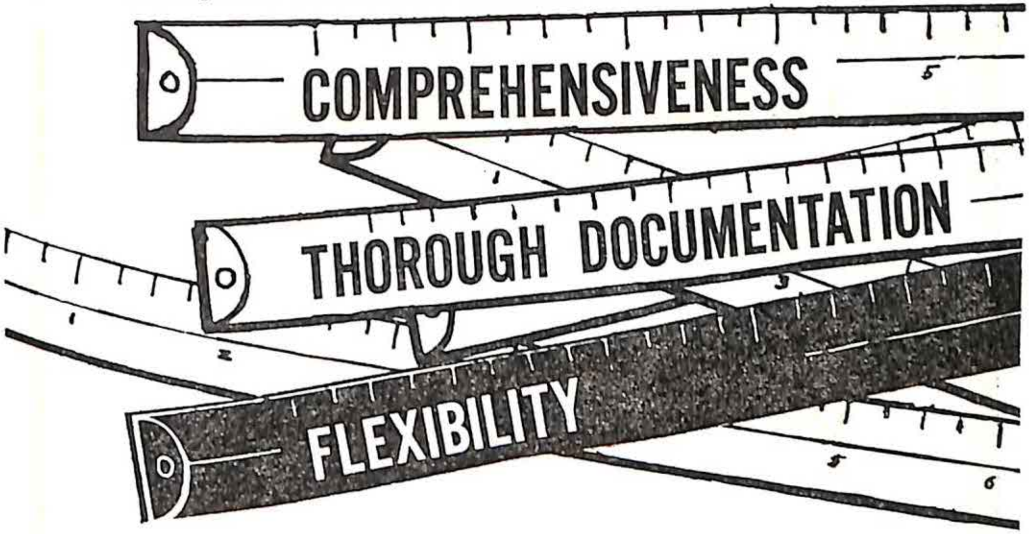
". . . both scholarly and lucid . . . discusses classic studies in the field, as well as the most recent research, within the context of contemporary classroom studies."
Barbara MacKenzie, *Brooklyn College of The City University of New York*

576 pages; 21 figures; 73 tables; $8.50

To accompany EDUCATIONAL PSYCHOLOGY; An INSTRUCTOR's MANUAL, prepared by the authors; 70 pages

For further information, write to
ALFRED A. KNOPF, *Publisher*
College Department 501 Madison Avenue New York 10022

How do you measure the value of a textbook?



Does the psychology text you are presently using measure up as well as this outstanding new textbook?

Heckel-Peacock

TEXTBOOK OF GENERAL PSYCHOLOGY

Designed for college and university courses in "General Psychology", this new text provides a broad overview of psychology at the introductory level. Comprehensive and thoroughly documented, its well-planned flexible format is sufficiently detailed to be used without supplementation for the usual one quarter or one semester course. Yet, it is flexible enough to be expanded to your specific requirements with reprints from the Bobbs-Merrill Series or from SCIENTIFIC AMERICAN.

Logically organized into six analytical sections, it introduces biological and developmental topics early in the presentation and emphasizes the developmental, physiological and historical areas. You will find that it provides a more expansive treatment of the *physiological*, *sensory*, *group* and *developmental areas* than other introductory textbooks. The easily understood discussions examine all essential aspects of modern psychology in depth with analogies to everyday behavior and experience.

A complimentary Teacher's Guide and Test Booklet is provided each instructor adopting this new book. Examine a copy of TEXTBOOK OF GENERAL PSYCHOLOGY and compare it to the text you are presently using. You may find that it contains *more* of the essential features you require for your introductory courses.

BY ROBERT V. HECKEL, Ph.D., Professor of Psychology and Director of Clinical Training, University of South Carolina, Columbia, South Carolina; and LELON J. PEACOCK, Ph.D., Professor of Psychology, University of Georgia, Athens, Georgia. Contributing author: ROBERT E. McCARTER, Ph.D., University of South Carolina, Columbia, South Carolina. Publication date: September, 1966. 480 pages plus I-XII. 7" x 10". Price, $8.50.

THE C. V. MOSBY COMPANY
3207 Washington Boulevard
MOSBY
Publishers
St. Louis, Mo. 63103

The American Psychological

Association **announces . . .**

a "Special Issue" of the

AMERICAN PSYCHOLOGIST

November 1966

The November issue of the *American Psychologist* will be a SPECIAL ISSUE entirely devoted to a consideration of the problems involved in the exchange of scientific and technological information within the "scientific community". The contributors are individuals who have distinguished themselves in such areas as psychology, sociology, physics and library science and who were selected so that the issue would present a variety of viewpoints in addition to those of the behavioral and social scientist. Among them are included Donald deB. Beaver, William D. Garvey, Belver C. Griffith, Don Marquis, Herbert Menzel, William J. Paisley, Edwin B. Parker, Derek J. de Solla Price, Philip J. Siegmann and Don R. Swanson.

Briefly, the articles will deal with topics relative to a) the use of data from the behavioral sciences in the improvement of techniques for the exchange of scientific information b) systems of information documentation, retrieval and transfer, and the impact and potential of automation in structuring of such systems and c) the relationship which should exist among science, technology and enterprise in order to derive maximum social utility from science.

The November SPECIAL ISSUE will be overprinted so that additional copies may be ordered from:

AMERICAN PSYCHOLOGICAL
ASSOCIATION, INC.
1200 Seventeenth Street, N.W.
Washington, D. C. 20036

Texts For Teachers. . .

Educational Psychology

Second Edition

By Glenn Myers Blair, R. Stewart Jones, and Ray H. Simpson, all of the University of Illinois

An unusually comprehensive and functional treatment of educational psychology, this book presents facts, principles, and methods of procedure, and reinforces them with illustrations from the psychology laboratory, clinical experience, and classroom experimentation. A teacher's manual is available.

1962, 678 pages, $7.95

Readings in Educational Psychology

Edited by Victor H. Noll and Rachel Noll, both of Michigan State University

Undergraduate and graduate students preparing to teach in the elementary and secondary schools will find in this book the broadest possible perspective in the field. The articles cover the latest methods, procedures, and classroom practices; results of new testing; evaluation and measurement research; and the most advanced findings in group dynamics, individual psychology, and mental health.

1963, 608 pages, $4.50

Statistics for the Classroom Teacher

A Self-Teaching Unit

By Edward Arthur Townsend, Westfield State College, and Paul J. Burke, The City College of The City University of New York

This supplementary text provides a review of the fundamentals of statistical concepts and procedures. The text explains the basic procedures for collecting and organizing data of the type most frequently encountered by the classroom teacher.

1963, 80 pages, $1.95

An Introduction to Educational Research

Second Edition

By Robert M. W. Travers, Western Michigan University

Recognizing the need for a modern account of research in educational development, the author fulfills a twofold purpose. First, he offers a basic guide for future educational researchers; and second, he provides an interpretation of the aims and methods of educational research for teachers, administrators, and students. The role of theory is emphasized throughout and distinguishes this text from others in the field.

1964, 608 pages, $7.50

Write to the faculty service desk for examination copies

THE MACMILLAN COMPANY

866 Third Avenue, New York 10022

A valid measure of student achievement in high school psychology

ENGLE PSYCHOLOGY TEST

Revised Edition

by T. H. Engle, Indiana University, Fort Wayne Campus

The newly revised *Engle Psychology Test* reflects the objectives, methods, and materials of introductory courses in general psychology now being taught to high school seniors throughout the country. It is also useful for appraising the background of students entering an introductory college course in psychology. The test evaluates the acquisition of factual knowledge and psychological principles and also understanding in dealing with real or hypothetical problem situations.

The manual that accompanies the test contains tables for interpreting raw scores in terms of percentile ranks or stanine ranges, based on a nation-wide sample of students in high school psychology courses.

Specimen sets are available to school administrators and supervisors; write to:

HARCOURT, BRACE & WORLD, Inc. • Test Dept.
757 Third Avenue, New York, N.Y. 10017
Chicago • Dallas • Atlanta • Burlingame

Film proves effective teaching aid
in courses in educational psychology
and introductory psychology . . .

BEHAVIOR THEORY IN PRACTICE

ELLEN P. REESE, *Mount Holyoke College*; filmed by Charter Oak Studios, New York. An hour and twenty minutes of 16 mm. color film with sound. Purchase price, complete film, $700.00, single reels, $200.00 each. Rental price, complete film, $45.00, single reels $15.00 each. Reel 3 for preview on request.

EDUCATIONAL PSYCHOLOGY, 2nd Edition

KARL C. GARRISON, Old Dominion College; ALBERT J. KINGSTON, *University of Georgia*; and ARTHUR S. MCDONALD, *Marquette University*. Extensively revised, the second edition of this text provides students of education with a sound background in the psychology of learning. 544 pp., illus., $6.50

APPLETON-CENTURY-CROFTS
Division of Meredith Publishing Company
440 PARK AVENUE SOUTH, NEW YORK, N. Y. 10016

Important Contributions to the Field of Educational Psychology

MEASUREMENT AND EVALUATION IN EDUCATION, PSYCHOLOGY AND GUIDANCE

GEORGIA SACHS ADAMS, Los Angeles State College, in consultation with THEODORE L. TORGERSON, Emeritus, University of Wisconsin

Here is a text that offers an uncommonly thorough treatment of measurement, both theory and practice, for students in elementary and secondary education, as well as psychology and guidance. Written and organized to meet the needs of both the average and the advanced student, it emphasizes understanding rather than computational skill.

1964 **667 pp.** **$9.50**

MODES OF THINKING IN YOUNG CHILDREN: A Study of the Creativity-Intelligence Distinction

MICHAEL A. WALLACH, Duke University
NATHAN KOGAN, Educational Testing Service

Sets forth the details of a methodologically rigorous investigation of creativity. The core of the authors' work is their isolation of an associational dimension of creativity that is independent of conventionally defined general intelligence. Essential reading for courses in which cognitive processes, personality, motivation, and social behavior in children are studied.

1965 **358 pp.** **$8.75**

Holt, Rinehart and Winston, Inc.
383 Madison Avenue,
New York, New York 10017

IMPORTANT NOTICE!

CHANGES IN THE APA JOURNAL PROGRAM FOR 1967

JOURNAL OF COUNSELING PSYCHOLOGY

Presently published quarterly by the Journal of Counseling Psychology, Inc. The APA will assume publication effective January, 1967 and the journal will be published *bimonthly*, one volume per year. Dr. Francis P. Robinson, Ohio State University, will continue as Editor. The regular subscription rate will remain $10.00 per year. (Subscriptions for 1966 should continue to be addressed to Journal of Counseling Psychology, Inc. Subscriptions for 1967 should be addressed to the American Psychological Association.)

PSYCHOLOGICAL MONOGRAPHS

Although Monographs will continue to be published, this will no longer be a distinct subscription publication. Based on the subject matter of a monograph, it will be considered a supplement to the related subject-area journal and distributed free as a separate to subscribers of that journal. In recent years about six monographs have been published annually in the subject area pertinent to the *Journal of Experimental Psychology*, six pertinent to the *Journal of Personality and Social Psychology*, two to the *Journal of Applied Psychology*, two to the *Journal of Comparative and Physiological Psychology*, and two to the *Journal of Consulting Psychology*. It is anticipated that a similar distribution of monographs will apply in future years, with other journals likely to publish monograph supplements from time to time.

JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY

As previously approved by the Council of Representatives, this journal will be published in three volumes per year, four issues per volume. The recommended page-allocation for 1967 is 1500 pages, an increase of 150 pages. In addition, it is expected that subscribers will receive as a part of their subscription approximately six monograph supplements. The regular subscription price will be $30.00 per year.

JOURNAL OF EXPERIMENTAL PSYCHOLOGY

As previously approved by the Council of Representatives, this journal will be published in three volumes per year, four issues per volume. The recommended page-allocation for 1967 is 2100 pages, an increase of 300 pages. In addition, it is expected that subscribers will receive approximately six monograph supplements as a part of their subscription. The regular subscription price will be $30.00 per year.

PSYCHOLOGICAL ABSTRACTS

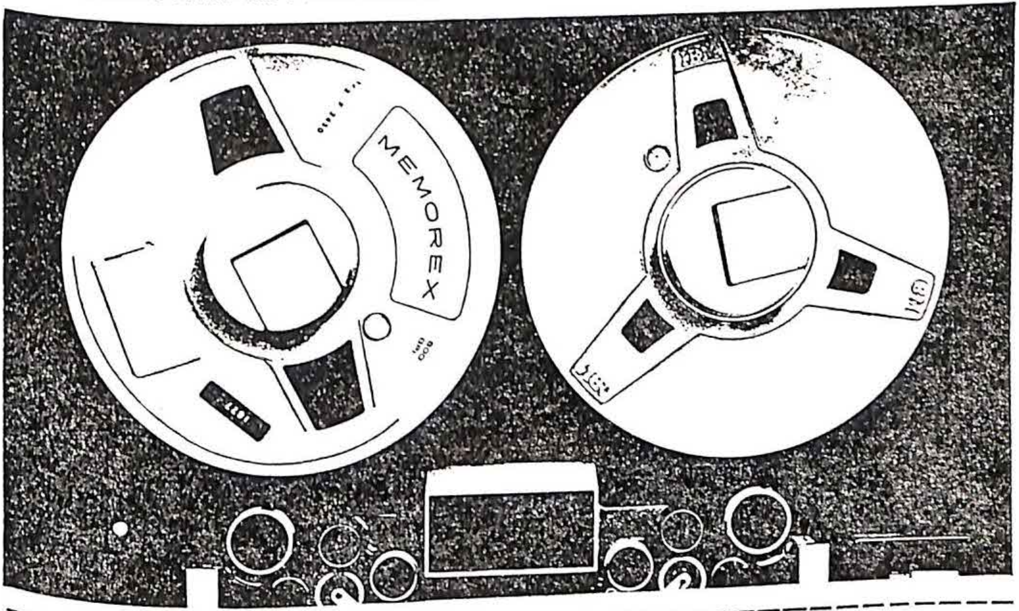
In the last three years the number of abstracts published annually in *Psychological Abstracts* has increased from approximately 8,500 to 17,000. This increase has been necessary in order to provide adequate coverage of the rapid growth of literature in the field of psychology. The increased coverage has been accompanied by increased cost of publication; consequently, the regular subscription price, as approved by the Council of Representatives, will be $30.00 per year.

SRA's reports supply data for your research

Do you want more than reports of scores and frequency distributions with your test results? Several of the reports available through our electronic scoring service contain highly technical information for your research.

For example, an Item Analysis furnishes data regarding group performance on each item in the test battery. Such analyses are available with the Achievement Series and the Iowa Tests of Educational Development. While the Achievement Series analysis supplies data pertaining to correct responses, the ITED analysis also includes information for each alternative.

Another report is the Research Plan. This new report for the ITED contains means, standard deviations, sums of squares and cross-products for each subtest in both converted scores and growth scale equivalents. In addition, the plan will provide correlations of each subtest with data entered in the Numeric Grid.



Many other technical reports are available. Clip this coupon for a brochure describing the various reports and our electronic scoring service.

Name__________

Position__________

School__________

Address__________

City__________ State_____ Zip_____

S R A

Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A Subsidiary of IBM

1966.

This has been a year of important contributions to the literature of educational psychology.

And Wiley is proud to have put a number of them into print.

PSYCHOLOGY in COMMUNITY SETTINGS: Clinical, Educational, Vocational, Social Aspects. *By* SEYMOUR SARASON, MURRAY LEVINE, I. IRA GOLDENBERG, DENNIS CHERLIN, *and* EDWARD BENNETT, *all of the Psycho-Educational Clinic, Yale University*. Describes the activities of the Psycho-Educational Clinic, which has effected marked departures in clinical training. 1966. 714 pages. $12.95.

SOCIAL CLASS and the URBAN SCHOOL: The Impact of Pupil BackGround on Teachers and Principals. *By* ROBERT E. HERRIOT, *Florida State University*; *and* NANCY ST. JOHN, *Harvard Graduate School of Education*. A study based on data from 490 schools in 41 large American cities. 1966. 289 pages. $7.95.

STUDYING the CHILD in SCHOOL. *By* IRA J. GORDON, *University of Florida*. A theoretical overview of classroom assessment techniques fresh from the educational and psychological laboratories. 1966. 145 pages. $4.95 in Cloth; $2.95 in Paper.

ROLE THEORY: Concepts and Research. *Edited by* BRUCE J. BIDDLE, *University of Missouri*; *and* EDWIN J. THOMAS, *The University of Michigan*. The first book devoted to the emerging specialization of role theory. 1966. 453 pages. $8.95.

INSTRUCTIONAL MEDIA and CREATIVITY: The Proceedings of the Sixth Utah Creativity Research Conference held at Torrey Pines Inn, La Jolla, California. *Edited by* CALVIN W. TAYLOR, *University of Utah*; *and* FRANK E. WILLIAMS, *Macalester College*. 1966. 397 pages. $8.95 in Cloth; $3.95 in Paper.

STUDIES in COGNITIVE GROWTH. *By* JEROME S. BRUNER, *Harvard University*; ROSE R. OLVER, *Amherst College*; PATRICIA M. GREENFIELD, *Research Fellow in Cognitive Studies*; et al. First major theoretical assessment of the process of cognitive development since the work of Jean Piaget and his colleagues. 1966. 343 pages. $7.95.

A THEORY of ACHIEVEMENT MOTIVATION. *Edited by* JOHN W. ATKINSON, *The University of Michigan*; *and* NORMAN T. FEATHER, *The University of New England, Australia*. A theory of achievement motivation that involves concepts of motive, expectance, and incentive value. 1966. 392 pages. $11.50.

PERSONALITY, SOCIAL CLASS and DELINQUENCY. *By* JOHN J. CONGER, *University of Colorado; and* WILBUR C. MILLER, *University of Denver*. A research project begun in 1956 for the investigation of teenage delinquency in Denver and suburban counties. 1966. 249 pages. $7.95.

Order from your Bookseller or

JOHN WILEY & SONS, Inc.
605 Third Avenue, New York, N. Y. 10016

New and Forthcoming from McGraw-Hill

CONDITIONING AND INSTRUMENTAL LEARNING

By *Wendell I. Smith* and *J. William Moore*; both of Bucknell University. 192 *pages*, $4.95 (*clothbound*), $2.50 (*soft-cover*). The first in a new series of programmed psychology books which provide extensive coverage of key topics taught in first-year psychology. Units can be assigned in lieu of text readings to reduce the need for classroom discussion.

ADOLESCENT DEVELOPMENT, Third Edition

By *Elizabeth B. Hurlock*, The Graduate School of Education, University of Pennsylvania. *McGraw-Hill Series in Psychology. Available Fall.* In an extraordinarily thorough and comprehensive treatment of its subject, Dr. Hurlock covers the traditional topics and extends the material to emphasize changing cultural patterns and the emergence of new problems for today's adolescent. An instructor's manual is available.

THE PSYCHOLOGY OF LEARNING, Third Edition

By *James Deese* and *Stewart H. Hulse*; both of the Johns Hopkins University. *McGraw-Hill Series in Psychology. Available Winter.* Provides as representative a picture as possible of the contemporary concern with the psychology of learning. Broad coverage and detailed exposition of basic problems are the most important features of the book.

SCALES FOR THE MEASUREMENT OF ATTITUDES

By *Marvin E. Shaw* and *Jack M. Wright*; both of the University of Florida. *McGraw-Hill Series in Psychology. Available Winter.* Here are all the worthwhile attitude scales in a single source. An extensive bibliography is included.

ELEMENTARY STATISTICS: Data Analysis for the Behavioral Sciences

By *Paul A. Games* and *George R. Klare*; both of Ohio University. *Available Winter.* Univariate descriptive statistics, statistical inference, bivariate descriptive statistics, and experimental design are all included in this treatment of major concepts for the beginning student. An instructor's manual is available.

BASIC MATHEMATICAL AND STATISTICAL TABLES FOR PSYCHOLOGY AND EDUCATION

By *William Meredith*, University of California, Berkeley. *Basic Tables Series. Off Press.* A collection of mathematical and statistical tables of value to students taking courses in psychology and education.

DICTIONARY/OUTLINE of Basic Statistics

By *John E. Freund*, Arizona State University; and *Frank J. Williams*, San Francisco State College. *Off press.* Includes a dictionary of statistical terms and an outline of statistical formulas. Basic statistical tables and a bibliography of statistics books are also included.

UNDERSTANDING EDUCATIONAL RESEARCH, Second Edition

By *D. B. Van Dalen*, University of California, Berkeley; and *William J. Meyer*, Syracuse University. *Psychology and Human Development in Education.* 544 *pages*, $7.95. Introduces the novice to the tactics and terminology of research work. Uses familiar classroom and life illustrations to acquaint the student with the logical, statistical, psychological, and theoretical foundations of scientific investigations.

Send for your examination copies today.

McGRAW-HILL BOOK COMPANY

330 West 42nd Street, New York, N. Y. 10036

Help him get ready for work

The new Rochester Occupational Reading Series offers both reading and language instruction, plus occupational information to poor readers who will soon be entering the job world.

By reading realistic stories and completing exercises that are mature in approach, learners gain a clear understanding of the skills and attitudes necessary for job success. They are led to develop such skills as completing application forms, reading comprehensively, performing simple arithmetical computations, and speaking simply and correctly. Basic social attitudes include: good manners, obedience to law, and good health practices.

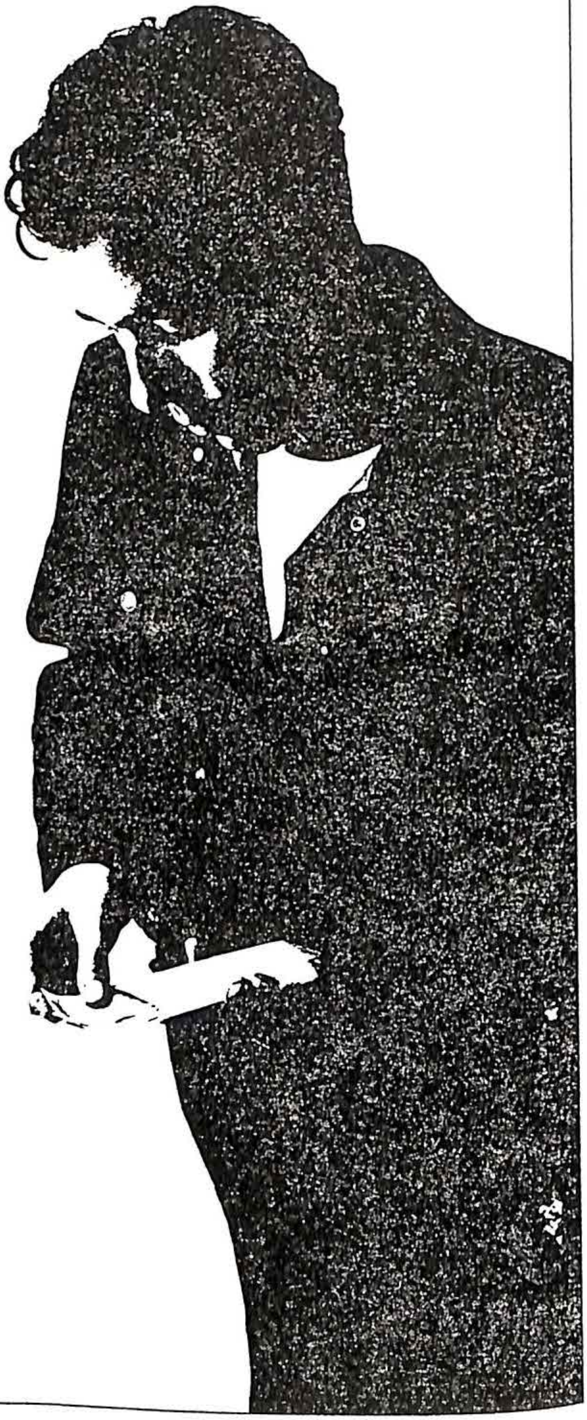
To use the program, a learner only needs second through fifth grade reading ability. All materials are written at three different levels within that range. They appear to be identical so older students working at the lowest level feel no embarrassment about their status.

Prepare your students for the occupational world. Write for a brochure describing this unique learning program — the new Rochester Occupational Reading Series.

SRA

Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A Subsidiary of IBM



Outstanding books for education courses

EDUCATIONAL PSYCHOLOGY
Second Edition
by Lee J. Cronbach

An extensive revision marked by a pronounced emphasis on intellectual learning and development and a full presentation of the psychological theory and research that underlie educational practice. ". . . the finest text in its field. It is readable but scholarly."—Paul B. Wood, Central Wesleyan College. Three separate Test Item Files available. 706 pages, $7.95

STUDENT GUIDE
for Cronbach's Educational Psychology
by Gabriel M. Della—Piana,
in consultation with Lee J. Cronbach

The purpose of the Guide is to help the teacher and prospective teacher apply the principles of educational psychology to concrete teaching situations. Organized to parallel Dr. Cronbach's text, each chapter in the Guide includes study questions, application exercises, and teaching unit exercises. Answers to the self-study questions and sample teaching units are contained in the Appendix section. Paperbound, 151 pages, $1.95

LEARNING
by J. Charles Jones

A concise, authoritative text that provides a selective introduction to the subject of learning. The author first considers education and behavioral change and then proceeds to a discussion of the characteristics of learning. Paperbound, 150 pages, $1.75 (probable). In preparation

MEASURING PUPIL ACHIEVEMENT AND APTITUDE
by C. M. Lindvall

A brief and lucid introduction to the principles of testing and evaluation that are essential to teachers for the effective assessment of pupil aptitude and achievement. The clarity of his presentation makes the book highly useful for in-service training programs or for teacher self-study. Paperbound, 128 pages, $1.50 (probable). Publication: January, 1967

Harcourt, Brace & World, Inc.
New York / Chicago / Burlingame / Atlanta

RUSSIAN MONOGRAPHS ON BRAIN AND BEHAVIOR

A series of translations of Russian Monographs concentrating upon the historicai background of investigations in neurophysiology and conditioned behavior as well as upon certain contemporary advances in these fields.

The translation program was initiated by H. W. Magoun and D. B. Lindsley and has been continued by the latter. The complete series will consist of eleven monographs published in hard-cover book form. To date, three of the eleven monographs have been printed and are available through the American Psychological Association.

Available now . . .

ESSAYS ON THE HISTORY OF PHYSIOLOGY IN RUSSIA

By Kh. S. Koshtoyants.

These essays review broadly the history of the development of physiology as a science in Russia, beginning in the early 18th century and continuing to modern times. They are more than a history for they deal with the personal characteristics and strivings of great scientists who forged teaching and research disciplines in anatomy, physiology, and psychology.

321 pp. $4.00

AUTOBIOGRAPHICAL NOTES

By I. M. Sechenov.

Sechenov, revered as the father of physiology in Russia, was born in 1829. His scientific and academic pursuits, though always stimulating and challenging, were fraught with difficulties due to political interference and harassment. These **Autobiographical Notes**, written about a year before his death in November, 1905, provide a vivid picture of his history.

174 pp. $4.00

ORIENTING REFLEX AND EXPLORATORY BEHAVIOR

Edited by L. G. Voronin, A. N. Leontiev, A. R. Luria, E. N. Sokolov, and D. S. Vinogradova.

A collection of papers read at a Conference organized in February, 1957 by the Institute of Defectology of the Academy of Pedagogical Sciences, RSFSR, and by the Department of Psychology and Higher Nervous Activity of the M. V. Lomonsov Moscow State University. As the varied experiments and points of view in this collection of papers will reveal, a serious attempt has been made to establish the role and significance of the orienting-exploratory reaction, not only in the elementary behavior of lower animal forms, but also in the complex behavior of higher forms, including man and his cognitive functions.

462 pages $4.00

AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 Seventeenth St., N. W., Washington, D. C., 20036

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 57 December 1966 Number 6

CONTENTS

Age and Mental Abilities: A Second Adult Follow-Up 311

W. A. OWENS

Structure of Intelligence in Negro and White Children 326

IRA J. SEMLER AND IRA ISCOE

Discriminability of Words, and Letter Cues Used in Learning to Read 337

S. J. SAMUELS AND W. E. JEFFREY

Six Structure-of-Intellect Hypotheses in Six-Year-Old Children 341

R. E. ORPET AND C. E. MEYERS

Methodological Note on Feeling-of-Knowing Experiments 347

J. T. HART

Effect of Selected Methodological Conditions on Dependent Measures Taken after Classroom Experimentation 350

WILLIAM L. GOODWIN

Modifiability of an Impulsive Tempo 359

JEROME KAGAN, LESLIE PEARSON, AND LOIS WELCH

Relations among Physical Skills and Personal and Social Variables in Three Cultures of Adolescent Girls 366

BOYD R. MCCANDLESS AND FAHMIDA ALI

Teachers' Belief Systems and Preschool Atmospheres 373

O. J. HARVEY, B. JACK WHITE, MISHA S. PRATHER, RICHARD D. ALTER, AND JAMES K. HOFFMEISTER

Cognitive Transfer in Verbal Learning 382

JAMES H. REYNOLDS

This is the last issue of Volume 57.

Volume Title Page and Contents of the Volume appear herein.

Journal of Educational Psychology
1966, Vol. 57, No. 6, 311-325

AGE AND MENTAL ABILITIES:
A SECOND ADULT FOLLOW-UP

W. A. OWENS[1]
Purdue University

The 96 Ss, first tested in 1919 and again in 1950 with the Army Alpha, were again tested with the same test in 1961 at an average age of 61 yrs. This paper reports the 1950–1961 results. The decade from age 50 through age 60 was one of relative constancy in mental ability test performance for the present Ss. The apparent trend, which is statistically insignificant, is slightly downward. Contrary to expectation, aging does not seem to have increased either individual or trait differences among the testees of 1961. Clearly, however, major components of the pattern of living of these Ss have served as moderators of the relationship between age and mental ability.

In 1953 the writer published a report of a longitudinal study of age changes in mental abilities. The subjects (*S*s) were 127 male freshmen of 1919 who had originally taken Army Alpha, Form 6, as an entrance examination at Iowa State University. They were retested during 1949–50 with the identical examination accompanied by a Personal Information Sheet composed of twenty "open-ended" items. Results obtained indicate a substantial and significant increment in test scores over the years, a finding subsequently confirmed in other longitudinal studies by Bayley (1955), Bentz (1953), Charles (1953), Nisbet (1957), and others. These results also suggest that individual differences and trait differences each remain of substantially constant magnitude; and that such variables as amount of education, field of college specialization, rural to urban migration, numbers of hobbies and recreational activities, and earned income are important correlates of temporal shifts in test score.

The initial study covered the *S*s' mean ages 19 to 50. The present follow-up extends this coverage approximately another decade to mean age 61. It was undertaken for at least two reasons. First, because the writer is unaware of comparable longitudinal, descriptive data; and, second, because it was believed that an attempt to identify the life-experience or life-history correlates of the anticipated shifts in test score would move beyond the simple and usual description of intellective changes towards an explanation of them.

Specifically, the problems of this investigation are three in number: (*a*) to describe the impact of an additional 10-year age increment upon the mental abilities measured by Army Alpha; (*b*) to appraise the effects of the given age increment upon the magnitudes of individual differences and trait differences, as measured; and (*c*) to reveal, via a scored life history, characteristics or patterns of living which are typical of those whose mental abilities age has treated more versus less kindly. In all cases, emphasis will be upon the last 10-year period, but an attempt will also be made to integrate the data of the last 10 years with that of the preceding 30.

---

[1] With the collaboration of Don C. Charles and John A. Bath and the assistance of Richard S. Gillmer, Paul R. Kelley, Michael Rosenfeld, and Lyle Schoenfeldt.

METHOD

In essence, these problems will be attacked in order by: (*a*) noting shifts, over time, in the levels of the mean verbal (V), numerical (N), and reasoning (R) components of Alpha or of its subtests, with some tentative correction for accompanying cultural changes; (*b*) comparing earlier and later variances, across tests and across people; and (*c*) item-analyzing an extensive scored life history against criteria of changes in the Alpha component scores from 1919 to 1950 and from 1950 to 1961, with subsequent factoring of the "valid" items.

Hypotheses

Based upon prior knowledge obtained from the first follow-up, upon considerable experience with life-history data and factors, and upon the substantial expansion in the research literature since 1953, a number of hypotheses were formulated prior to the analysis of the present data.

The first and most general of these is that there *would* be significant numbers of life-history item responses significantly related to magnitude of shift in component Alpha scores over time. Surely the work of Kallman and Jarvik (1959) on senescent twins implies the genetic conditioning of a decline in many critical characteristics. On the other hand, the work of Hebb (1949) on early deprivation might be interpreted to suggest that adult intelligence varies with the quantity and quality of past experience, a point of view espoused and elaborated by Hunt (1961, p. 363). If this latter position is tenable, and the present data permit an adequate test, then the preceding hypothesis should be sustained.

The second hypothesis is that the two point estimates of ability previously obtained (1919 and 1950) lie along a typical parabolic growth curve now extended parallel to the base line; thus, that there should be no significant change in overall level of score from 1950 to 1961. This position is in contrast to the interpretation of Jones (1959), who believed that the present *S*s had probably improved their potential scores so greatly during the college years that a subsequent decline was still more than compensated. If Jones' view were correct, the decline begun at the *S*s' mean age 23 should continue and accelerate through their mean ages 50 to 60. Conversely, if no decline is apparent during this last decade, it seems highly unlikely that one would have been detected some 38 years earlier.

The third hypothesis is that, based on an assumption of the onset of some differential organic changes, slight but satistically significant increases in the magnitudes of both individual differences and trait differences from 1950 to 1961 should be revealed.

The fourth hypothesis, assuming a continuous nature-nurture interaction, is that socioeconomic status should be a correlate of relative improvement in test performance.

The fifth hypothesis is that the long-range importance of motivational factors is so great that a number of what Wechsler (1950) might call "nonintellective determiners of intelligence" should be found to be related to a relative improvement in test performance.

The sixth hypothesis is that by virtue of some recent organic changes in some *S*s, failures in health and vigor should be found to be related to a relative decrement in test performance from 1950 to 1961.

The seventh hypothesis is that the social introvert, less accustomed to social interaction, would be less disrupted by some essential age-related disengagement: thus, that introversion should be positively related to test performance over the interval 1950–1961.

Subjects

The *S*s were 96 males, of mean age 61 in 1961, who had originally been tested at Iowa State University as entering freshmen in 1919. The typical *S* may be characterized as follows: (*a*) He lived in a rural area or small village prior to college but has since moved to the city; (*b*) he graduated from college with a major in agriculture or engineering; (*c*) he is married, owns his own home or apartment and has had one or more children; (*d*) he still resides in the Middle West; (*e*) his occupation is found toward the upper end of the occupational-intelligence hierarchy; and (*f*) his reported net income for 1960 was between $7,000 and $13,000.

The 96 *S*s on whom data are here available represent a 75% sample of the 127 first retested during 1950. Of the 31 lost, 13 were deceased, 5 were disabled, 5 could not be located, and 8 refused to participate. (A further note regarding bias appears later).

Measuring Instruments

Two measuring devices were utilized in this study. The first is the Army Alpha Examination, Form 6. It requires approximately 25 minutes of testing time and Table

TABLE 1
SUBTEST AND COMPONENT RELIABILITIES OF ARMY ALPHA, FORM 6

Subtest or component	Retest reliability			Component predicted
	1919 versus 1950	1950 versus 1961	1919 versus 1961	
1. Following directions	.30	.46	.41	Reasoning (R)
2. Arithmetical problems	.69	.83	.68	Numerical (N)
3. Practical judgment	.56	.68	.60	Verbal (V)
4. Synonym-antonym	.64	.71	.56	Verbal (V)
5. Disarranged sentences	.48	.69	.52	Verbal (V)
6. Number series completion	.62	.61	.54	Verbal (V)
7. Analogies	.56	.83	.54	Reasoning (R)
8. Information	.63	.75	.59	Verbal (V)
9. Verbal component	.75	.87	.73	—
10. Numerical component	.71	.86	.73	—
11. Reasoning component	.57	.85	.58	—
12. Total	.79	.92	.78	—

1 contains data relative to its reliability, components, and the subtest predictors of these components.

The second measuring device employed was a so-called "Life Experience Inventory" (L.E.I.) especially constructed for project purposes. It contains 228 objective items, of which 115 appear to have their options on a continuum and 113 do not. Those chosen for inclusion were selected from a file of perhaps 1,000 items on the basis that (*a*) their apparent content was such as to imply their relevance to intellectual change, and (*b*) if previously tried, they had validated against a seemingly relevant criterion or factor.

Data Collection

All *S*s were first invited to attend a class reunion and banquet held in conjunction with the project. Of the 96 ultimate participants, 28 attended this reunion, at which ½ day of their time was used for test and questionnaire administration. Room and board were provided for these *S*s, and they were made a token payment of $5.00 for services rendered.

The *S*s unable to attend the reunion were retested by examiners resident in their vicinity. These *S*s received $10.00 for participation, and their examiners received $10.00 per *S* and travel expenses for their services. Most examiners were members of the American Psychological Association, all were competent to administer group tests, and each was provided with detailed instructions for the administration of Alpha and the Life Experience Inventory.

Component Scores

In the interest of parsimony, intelligibility, and the enhancement of generalization, shifts in Alpha scores are reported in component score form wherever appropriate. As a consequence, life-history items potentially diagnostic of shifts in intellective abilities were selected for their demonstrated relationship to a temporal shift in one or more of these component scores.

As previously described (Owens, 1953), all raw scores on each of the eight subtests of Alpha were referred, prior to any computation, to appropriate norm distributions for 1,000 cases and were "read out" as normalized standard scores. To combine these eight transformed subtest scores into component scores, Guilford's (1954) report of his factorial analysis of Alpha was utilized. In it he employs multiple regression to weight three independent sets of subtests in such fashion as to optimize their relationship with the factor on which they originally loaded most highly. Since Guilford's weights were in raw score form, they were multiplied by the standard deviation of the norm group in each case to convert them to standard score weights. The minimum weight, within a factor, was then arbitrarily set equal to unity, although the ratio of weights within factors was, of course, maintained. Employing these effective weights verbal (V), numerical (N), and reasoning (R), component scores were derived from the given subtest scores for each of the three administrations of Alpha (1919, 1950, and 1961).

No criteria, external to Alpha, were available in terms of which to document the reality of the apparent intellective changes implied by shifts in test scores. However, shifts in the component Alpha scores themselves provided the criteria against which the items of the Life Experience Inventory were validated. Thus, a hypothetical change could occur in the V, N, or R component or in the total score; it could also occur over either the first time period (1919–1950) or the second (1950–1961).

An approach of this sort implied at least two questions. The first is—are several criteria, all derived from scores on the same

TABLE 2

CRITERION INTERCORRELATIONS FOR TWO TIME PERIODS

Component	1	2	3	4	5	6	7	8
1. Verbal, 1950–1919[a]		−.16	.19	.76	−.23	.07	−.12	−.25
2. Numerical, 1950–1919			.18	.17	.24	−.47	−.16	−.04
3. Reasoning, 1950–1919				.71	.18	−.10	−.42	−.12
4. Total, 1950–1919					.01	−.12	−.35	−.25
5. Verbal, 1961–1950						.00	.01	.72
6. Numerical, 1961–1950							.21	.36
7. Reasoning, 1961–1950								.56
8. Total, 1961–1950								

[a] The "Verbal, 1950–1919" criterion is the 1950 standard score minus the 1919 standard score, across all *Ss*.

test, reasonably independent? The relevant criterion intercorrelations appear in Table 2. Excepting subtest versus total score relationships of some spuriousness, it will be noted that considerable independence is, indeed, the rule.

The second question concerns the reliability of the component difference scores. This was crudely estimated for the V component, which is least speeded, by correlating the sum of the difference scores for Subtests 3 and 5 with the sum for 4 and 8. This procedure approximately equates factor loadings in the two "halves" of the component. Since two temporal intervals were involved (1919 to 1950 and 1950 to 1961), two coefficients were obtained and corrected for length through application of the Spearman-Brown formula. The corrected estimates were .37 for the first time period (1919–1950) and .51 for the second (1950–1961).

A decision to utilize the difference score approach in spite of these relatively low reliabilities resulted from several considerations more or less as follows:

(1) A difference score is the most direct index to that which it is desired to measure—changes in test score over time.

(2) Heterogeneity of subtest content argues that both the .37 and the .51 quoted are likely to be underestimates of the intrinsic reliability. For example, the median intercorrelation of Subtests 3, 4, 5, and 8 in 1919 was only .50, and the coefficients ranged from .43 to .52.

(3) The test-retest intervals involved are very large, particularly the first one; so that a great deal of true change in *S*s, differential by subtest, is undoubtedly assigned to error variance or unreliability.

(4) Only group differences are of present concern; no individual diagnosis is involved.

(5) Unreliability in the difference scores, as criteria for the analysis of life-history items, should cause conservative rather than radical errors.

Statistical Analysis

All computations involving the test scores were performed on the normalized standard scores for the subtests and total or on the component scores derived from them. The norm distributions of 1,000 comparable cases were arbitrarily assigned a mean of 5 and a sigma of 1. Shifts in mean score are, thus, reported in units of the norm standard deviation. Significance of these shifts was evaluated through application of the *t* test to a given distribution of difference scores to determine whether or not its mean departed significantly from that of a distribution with mean zero.

The significance of the affect of the given age increment on the magnitude of individual differences in a given function was evaluated by obtaining a 1950 and a 1961 variance, and by testing for the significance of the difference between these correlated variances in the fashion jointly described by Pittman (1939) and Morgan (1939).

The comparable effect of age on the magnitude of trait or "profile" differences was determined by computing two "V" scores for each *S*, one for each testing, each equal to the variance of his eight subtest standard scores about their own mean. A distribution of the differences between these two series of V scores was then obtained and its mean tested for significance in the usual way.

An analysis of the life-history correlates of the shifts in Alpha scores over time was made by relating response on each item of the L.E.I., successively, to each of the eight criterion segments (V, N, R, or T for either 1919–1950 or 1950–1961). In the case of the 113 noncontinuum items each option was regarded as a dichotomy, checked or unchecked; the given criterion was also dichotomized, and a direct binomial test of significance was applied (Siegel, 1956). However, no further attention will be herein devoted to the noncontinuum items, since the number of them which discriminated could easily have been accounted for as a chance event.

In the case of the continuum items, product-moment correlations were run between item score and each of the eight criterion scores. Those items which did not correlate with two criteria above the 20% probability level or with one criterion above the 10% level were eliminated. Brozek and Tiede's (1952) test was then applied to determine the probability of obtaining the given number of significant items by chance.

A clear limitation of the item analysis phase of this study is that the number of the *S*s involved ($N = 96$) is so small as to virtually preclude meaningful cross-validation. However, since the number is fixed, the practical issue seems to be one of how to obtain a maximum of information, and a minimum of misinformation, from the data available.

The following precautions were taken in the interest of minimizing the role of chance in the obtained results. (*a*) Prior hypotheses regarding the outcomes to be anticipated were set forth, and positive interpretations are restricted to the coincidence of results with expectation. (*b*) Since item interpretations are tenuous at best, and since the hypotheses offered implied the necessity, the discriminating items were factored to obtain a more parsimonious and generalized interpretation. (*c*) Brozek and Tiede's (1952) test was applied to assure that the number of discriminating items identified was *larger* than could reasonably be attributed to chance. Given this, it was felt that the interpretation of factors would not be greatly affected by the presence of an occasional "chance-discriminator" item. (*d*) As noted in the preceding paragraph, standards set for initial item selecting were relatively high.

The factorial method was applied to a 90 × 90 matrix, the elements of which were the 82 discriminating continuum type items of the L.E.I. plus the eight criterion difference scores of Alpha. Communalities were estimated by the image-covariance method, and 20 factors were extracted according to the principal axis method (Harman, 1960). Latent roots for these 20 factors were plotted for inspection and both five and nine factors were rotated to oblique simple structure employing the biquartimin criterion (Carroll, 1957). A panel of eight judges was then utilized to aid in interpretation. Three judges saw only Rotation A, three received only Rotation B, and the remaining two judges saw both. Decisions regarding interpretability and naming were based on a consensus of the judges.

RESULTS

Level of Abilities

The first major finding of this investigation is that the Alpha scores of these 96 *S*s remained remarkably constant over the period from 1950 to 1961.[2] Table 3 contains the mean differences and associated t values. It will be noted that at the subtest level none of the differences is signifi-

TABLE 3

TEMPORAL SHIFTS IN SUBTEST, COMPONENT, AND TOTAL ALPHA SCORES

Subtest or component	$\bar{x}$ 1961 − $\bar{x}$ 1950	t
1. Following directions	.001	.012
2. Arithmetical problems	−.237	−.436
3. Practical judgment	−.118	−.192
4. Synonym-antonym	−.049	−.072
5. Disarranged sentences	−.036	−.052
6. Number series completion	−.084	−.116
7. Analogies	.013	.025
8. Information	.043	.089
Verbal component	−.032	−.90
Numerical component	−.203	−4.38**
Reasoning component	.015	.28
Total	−.047	−.138

Note.—$M = 5$; $SD = 1$.
** = $p < .01$.

---

[2] Means of those tested in individual versus group situations did not differ significantly, although the former appeared slightly but consistently higher in 1961.

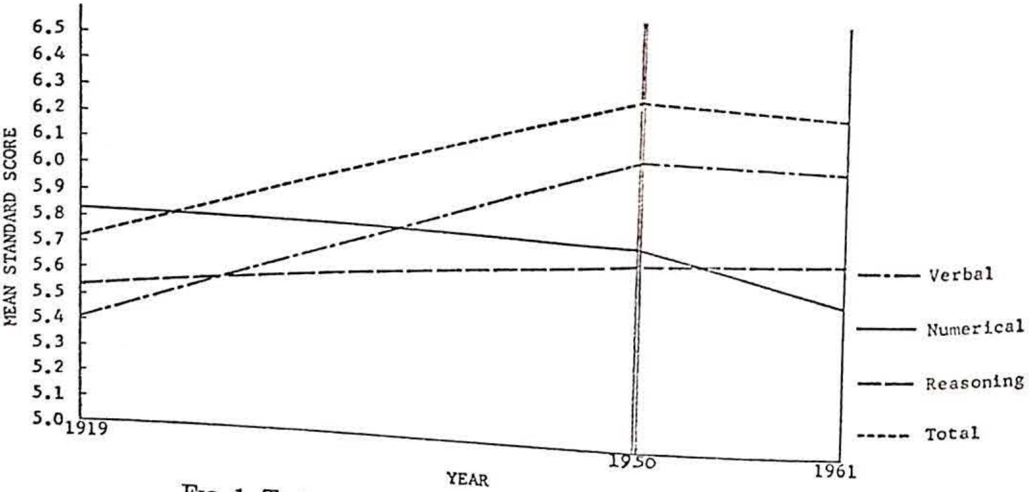


FIG. 1. Test performance trends for components and total.

cant; indeed, of the apparent "gains," five are negative and three positive. In the case of the components, the largest and third largest subtest losses have combined to produce a significant downward shift in the N component. On the total score, however, the t value does not closely approach significance and the apparent loss is of the order of magnitude of only one-tenth of a standard deviation. Figure 1 shows this trend in component scores and the total over the 42-year period of the three testings. Since the true functions cannot be determined from three point estimates, it seemed most parsimonious to connect these points with straight lines. However, the argument from the data is clear: *Accelerating declines in these abilities begun 30 to 40 years ago should be more apparent in the last decade than these declines are.* The definite implication is that any declines which have begun have done so more recently. A possible exception is the case of the N component. Here, indeed, potential test performance may have begun to slip at some point within perhaps 10 years of college graduation. The speculation is lent credibility by the fact that the sample is heavily loaded with engineers who had no doubt reached some peak of quantitative sophistication at college graduation; and by the knowledge that many of these *Ss* would shortly have moved into administration, into less technical aspects of engineering, or out of the field entirely. Any such move would clearly lessen their need for, and ability to maintain, quantitative knowledges and skills.

The second hypothesis stated was that there would be no significant change in overall level of score from 1950 to 1961. This hypothesis is sustained.

Individual and Trait Differences

Data pertaining to the third hypothesis stated appear in Tables 4 and 5. In the former, a comparison of the variances for 1950 and 1961 reveals that four differences are apparently negative and four positive and that none is statistically significant. It is therefore implied that no significant shift in the magnitude of individual differences took place over the 11-year period in question.

Table 5 contains comparable evidence regarding the effects of the

TABLE 4

TEMPORAL CHANGES IN INDIVIDUAL DIFFERENCES, 1950–1961

Subtest	1950 σ^2	1961 σ^2	±[a]
1. Following directions	.860	.742	−.805
2. Arithmetical problems	.928	.778	−1.538
3. Practical judgment	.578	.609	.347
4. Synonym-antonym	.683	.900	1.908
5. Disarranged sentences	.656	.844	1.694
6. Number series completion	.701	.650	−.464
7. Analogies	.812	.682	−1.508
8. Information	.425	.517	1.444
Total Scores	.700	.767	1.138

[a] None of these values is significant.

given age increment upon trait differences. Significance is not approached, and, although this sample contains a larger number of positive shifts, the negative sum slightly exceeds the positive. In accordance with this evidence, Hypothesis 3, anticipating an increase in both individual and trait differences, must be rejected.

TABLE 5

TEMPORAL CHANGES IN TRAIT DIFFERENCES, 1950–1961

Subtests	(σ^2 1961 − σ^2 1950)	Range	$N\pm$	t[a]
1 through 8	−3.2278	+1.856 −2.8870	+52 −44	−.602

Note.—$\sigma^2$1961 = the variance of the individual's subtest scores in 1961. Range = greatest individual increase and decrease in standard deviation units. $N\pm$ = number of cases showing an increase or decrease.

[a] Value is not significant.

L. E. I. Correlates

With respect to the third major problem stated, 82 of the 115 continuum-type life-history items utilized were found to be related to shifts in test score over time. If the assumption is made that the criteria are perfectly correlated, so that there is really only one criterion, then Brozek and Tiede's test (1952) indicates that the probability of obtaining 82 significant items from a pool of 115 by pure chance is far less than .0001. However, the criterion intercorrelations of Table 2 show that the "one criterion" assumption is unlikely. Thus, a more conservative and realistic probability estimate may be obtained by assuming complete independence of the eight criteria. If this is done, then the total number of possible item-criterion correlations is 920, to be compared with an obtained number of significant item-criterion correlations of 214. Given these figures, and the 10% probability level as the mean item discrimination level, this outcome should occur by chance far less often than one time in ten thousand. It thus seems safe to conclude that the first, general hypothesis of this investigation is tenable; that is, there *are* significant numbers of life-history correlates of shifts in Alpha score over time.

Factor Analysis

Since no analysis of the noncontinuum items of the L.E.I. was indicated, it seems important to note that the 82 continuum items gave very good coverage of the item categories originally established. As previously noted, these 82 life-history items, plus the eight criteria, constituted the entries in a 90 × 90 correlation matrix to be factored. Concurrent with the extraction of 20 factors from this matrix, Figure 2 contains a plot of the latent roots on which the apparent inflection point has been indicated. A decision to rotate both five and nine factors followed from an examination of this plot. Of the two, the five-factor rotation proved much more interpretable to the judges. The three who attempted to interpret the nine-factor rotation showed little consensus, particularly for the last four factors, and complained of ambiguity

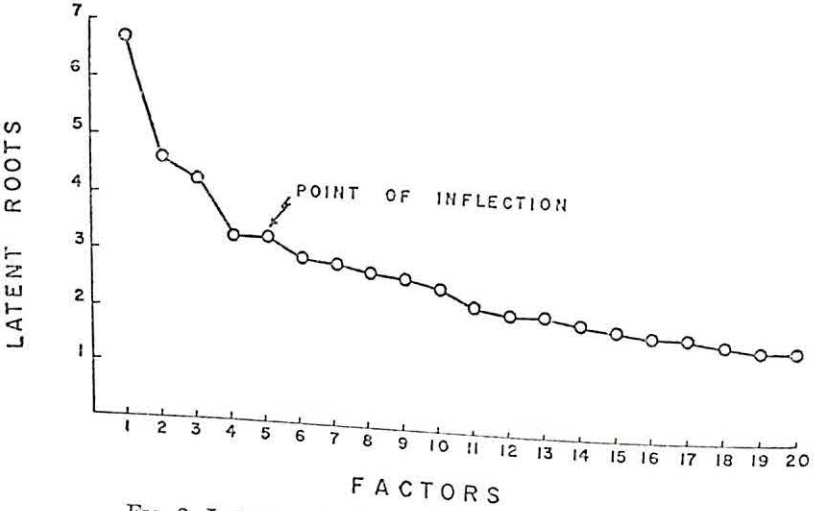


FIG. 2. Latent roots for 20 factors prior to rotation.

and overlap among them. On the other hand, the three judges who saw only the five-factor rotation, and the two who saw both, agreed very well in their interpretations. When coupled with the fact that very few item loadings dropped appreciably or disappeared in moving from nine factors to five, this evidence led to the adoption of the five-factor rotation. Although it was, nominally, an oblique solution, with a biquartimin criterion of simple structure, the intercorrelations among the factors were so low as to render orthogonal and oblique solutions virtually indistinguishable. Factor intercorrelations actually ranged from .11 to −.09 with a median midway between .04 and −.05.

TABLE 6

NUMBER OF VARIABLES LOADING ON GIVEN FACTORS

Factor	Items loading .30 or greater	Criteria loading .20 or greater
I	17	2
II	13	4
III	14	1
IV	15	1
V	11	3

The number of items with loadings of .30 or greater on a given factor, and the number of criteria with loadings of .20 or greater, appear in Table 6. It should particularly be observed that although four criteria load above .20 on Factor II and only one on Factors III and IV, still, each factor has at least one loading of the given magnitude.

In what follows a consensus of five judges who reviewed these factorial results has been employed in the interpretation and naming of the factors. The sense of each item has been briefly indicated, and the items listed in the order in which they load on the factors. The approximate percentage of the common factor variance accounted for by each factor has also been noted.[3]

Factor I. Socioeconomic Success was the name assigned to the first and largest factor, which accounted for approximately 12% of the common factor variance. The item response pattern implies the following characterization of high scorers on the factor.

---

[3] Percentages of common factor variance accounted for should be regarded as approximate, since the solution was not, formally, orthogonal.

Characterization of Ss	*Loading*
1. Rate high occupationally	.77
2. Hold a large amount of life insurance	.75
3. Receive a large income	.74
4. Have moved up occupationally	.67
5. Live in a large city	.63
6. Supervise people	.57
7. Prefer to have many projects "on the fire" simultaneously	.51
8. Save a large part of present income	.41
9. Prefer to work with others, rather than alone	.39
10. Work faster than most	.39
11. Goals have been fulfilled or surpassed	.37
12. Dislike direct supervision	.35
13. Married early	.35
14. College contributed to life's goals	.34
15. Receive ample recognition for a good job	.32
16. Are disturbed by leaving something unfinished	.31
17. Parents granted them great independence in high school	.31

The characteristic most clearly implied seems to be a high level of occupational and financial success. There are also understandable overtones of what might be called "executive attitude" exemplified in a desire for minimal external control, acceptance of responsibility, and a high drive level at work.

Factor II. This second factor, which accounts for approximately 8+% of the common factor variance, was designated as Sensitive Intelligence. Again, the item response pattern implies the following characterization of high scorers on the factor.

Characterization of Ss	*Loading*
1. Upset by some tragedy in youth	.51
2. Recognize a number of Alpha items from past administrations	.48
3. Frequently write down ideas	.42
4. Tell jokes frequently	.42
5. Devote much time to reading	.41
6. Are sometimes tense and nervous	.40
7. Keep some records and budget expenses	.39
8. Are quite religious	.37
9. Bring problems from home to work	.35
10. Have long job tenure	.33
11. Often feel dissatisfied with themselves	.33
12. Regard recent past as happiest period of life	.31
13. Health has been a social handicap	.30

A hypothetical individual making these responses evidences above-average intelligence in his devotion to reading, his superior memory, and his production of ideas judged worth recording. Coupled with these, however, is a syndrome of emotionality and sensitivity which may be inferred from a stated youthful tragedy, some tenseness and nervousness, some dissatisfaction with self, and some perseveration with problems.

Factor III. The judges described and labeled this factor as Physical Vigor. It accounts for a little less than 8% of the common factor variance. Items which load heavily on the factor imply the following characterization of a high-scoring respondent.

Characterization of Ss	*Loading*
1. Describe their physical vigor as above average	.70
2. Say work is not tiring	.49
3. A typical day does not tire them out	.47
4. Have taken advantage of their opportunities	.46
5. Concentrate on the technical aspects of the music at a concert	.43
6. Rarely feel dissatisfied with themselves	.43
7. Feel more than adequate for their jobs	.40
8. Say health has rarely, if ever, kept them out of activities	.38
9. Consider themselves more successful than average in things attempted	.37

Characterization of Ss	*Loading*
10. Feel they worked up to potential in college	.37
11. Have rarely had an idea they had to write down before forgetting it	.37
12. Formed occupational goals early in life	.35
13. Feel they don't receive adequate recognition for a good job	.33
14. Earned most of their college expenses	.31

Here, stated physical vigor, work attitudes, self-perceptions, and health history all suggest the appropriateness of the assigned name.

Factor IV. Introversion was the label assigned to this fourth factor, which accounted for about 6% of the common factor variance. The following response pattern defines the factor.

Characterization of Ss	*Loading*
1. Do not particularly enjoy social gatherings	.50
2. Do not get together socially with business associates	.47
3. Feel people are not all good	.45
4. Lived in a town or city, rather than a rural area, prior to college	.43
5. Consider themselves fairly tense or nervous	.43
6. Mother died at an advanced age or is still living	.42
7. Prefer to work on one thing at a time	.41
8. Have experienced major surgery more times than the average *S*	.38
9. Health has sometimes inhibited social activities	.38
10. Are considerably disturbed if something is left unfinished	.34
11. Usually ignore treatment for colds	.33
12. Have turned down social engagements because they were tired	.32
13. Activities have been suspended because of accident more often than average *S*	.32
14. Are reluctant to express their views in a group situation	.31
15. Have become less religious with age	.30

The name assigned to this factor seems clearly relevant to the implied lack of social enjoyment and activity, somewhat negative attitude towards others, possibly overlong association with mother, and suggestions of poor health, low energy level, and social reticence.

Factor V. The fifth and final factor was called Egocentric Independence, and it accounted for a bit less than 6% of the common factor variance. The characterization implied by the items of this factor is as follows.

Characterization of Ss	*Loading*
1. Consider their own feelings before those of others	.54
2. Not engaged in social welfare activities	.46
3. Prefer to work alone	.44
4. Are less religious than the average *S*	.42
5. Frequently spend an evening at home in leisurely reading	.41
6. Consider their reading speed to be above average	.41
7. Seldom make specific plans regarding the future	.38
8. Require eight or more hours of sleep	.34
9. Do not feel their fathers were particularly successful	.33
10. Life's work has fulfilled or surpassed goals	.33
11. Have never found it necessary to keep exact financial records or budgets	.31

A very high scorer on the factor would appear to be characterized by some selfishness and solitariness, by a preference for working alone, by an absence of religious belief, and by a feeling that he has done well, probably better than his father and better than he himself expected.

Factor-criterion relationships. Table 7 contains the critical evidence regarding relationships of Alpha component gains, representing intellective changes, to life-history factors, representing demographic characteristics and experiences. Given these data, it is now in order to recur to the

TABLE 7

LOADINGS >.20[a] OF ALPHA COMPONENT GAINS ON LIFE-HISTORY FACTORS

Life-history factors	Verbal component		Numerical component		Reasoning component		Total score	
	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
I. Socioeconomic Success	.52				.28		.23	
II. Sensitive intelligence					.31	.22	.69	
III. Physical vigor			.35					
IV. Introversion								.29
V. Egocentric independence		−.50	.36	.25				

Note.—P_1 = 1919–1950; P_2 = 1950–1961.

[a] This level was chosen because of the priorly noted unreliability of the component gains.

remaining four of the hypotheses originally enunciated.

The first of these asserted that socioeconomic status would be a correlate of intellective change. Table 7 shows that V and R component difference scores and the total score load .52, .28, and .23, respectively, on the factor Socioeconomic Success during the first time period (1919–1950). Hypothesis 4 may thus be regarded as confirmed.

The fifth hypothesis specified that motivational factors would be found to be related to intellective change. It will be observed in Table 7 that the R component difference score and the total both load substantially on the factor Sensitive Intelligence, the former for both time periods (1919–1950 and 1950–1961). To the extent that this factor involves early trauma, tenseness and nervousness, perseveration with problems, and some dissatisfaction with self, it has a motivational component. However, since the factor is considerably broader than this, Hypothesis 5 may be safely regarded as only tentatively confirmed.

The sixth hypothesis was that organically conditioned decrements in health and vigor would be found to be related to a relative decrement in test performance from 1950 to 1961. It may be noted in Table 7 that the numerical component difference score did load .35 on the factor Physical Vigor, but that it was for the first time period (1919–1950) rather than the second. The hypothesis, as stated, is thus rejected, although the observed relationship is of interest in its own right and the subject of a subsequent speculation.

The seventh hypothesis was that the social introvert would be less disturbed than his fellows by some essential age-related disengagements, apparent during the second time period, and should thus perform relatively better on Alpha. Here Table 7 does, indeed, reveal that the total difference score loads .29 on the factor of Introversion. The hypothesis, nevertheless, should probably be regarded as in some doubt for two reasons. First, the loading of .29 is not large; and, second, it stands alone. At best, only a very tentative sort of confirmation may be indicated.

Finally, it should be pointed out that the identification of a factor such as Egocentric Independence was not anticipated. Since there are three fairly substantial loadings on it, and no prior hypothesis, some comments regarding the possible significance of the relationships involved appear in the Discussion section which follows.

DISCUSSION

Interpretations

Some implications of the data presented are incompletely documented but of possible significance and concern.

Some years ago Kuhlen (1940) pointed out that cultural changes might well be mistaken for age changes in the context of a cross-sectional study. Clearly, the longitudinal study may also be distorted by cultural factors. Any projection of the intellective changes in 50-year-olds from 1950 to 1960 into those to be expected of comparable age groups from 1960 to 1970 must obviously be conditioned upon the assumption of comparable cultural evolution over the latter time period. However, over any long term such an assumption is clearly unrealistic. In such a case, it is probably most useful to revise the implied question so that we now inquire what a given age increment may be expected to do to the mental abilities of individuals in a *constant culture*. Cast in these terms, it becomes of interest to attempt to estimate "true" or intrinsic changes in the individual, separate and apart from the accompanying changes in the culture. One of Campbell's (1963) quasi-experimental designs was adapted to this purpose.[4] It requires that one "baseline" the apparent age changes noted by measuring a sample of present 19-year-olds comparable to the *S*s of this study at their age 19. Assuming comparable individuals, one may then attribute group differences observed to cultural change. If these differences are subtracted from apparent or observed changes in the same individuals over time, the residuals should represent "true" or intrinsic changes in these individuals. To fulfill the indicated purpose Alpha, Form 6, was administered during 1961–62 to a random sample of 101 male freshmen at the same institution attended by *S*s of the longitudinal study, that is, Iowa State University.

Table 8 contains the relevant data. Column 2 shows total "apparent" gains from 1919 to 1961. Column 3 contains as "cultural changes" the scores of the 1961–62 freshmen minus those of the freshmen of 1919. Under "true gain," by subtest and component, appear the differences between Columns 3 and 4. These corrected component gains compare with the uncorrected total gains of approximately $.5\sigma$ on V, $-.3\sigma$ on N and $.2\sigma$ on R shown in Figure 1. It is, thus, in the case of R that the correction introduces the most dramatic change. The explanation may well be fairly simple. The Analogies test, for example, was probably quite unfamiliar to freshmen of 1919, as were all group tests of intelligence. In 1961 it was not unfamiliar to either 19-year-olds or 60-year-olds. Thus, what appeared as a small intrinsic gain in R for the 61-year-olds may be better accounted for in terms of a vastly increased general level of sophistication with this type of test, in the midst of which the senior *S*s really lost ground in the relative sense.

In any event, the corrected components seem to be in better accord with both theory and practice than the uncorrected. For example, deterioration is commonly measured in terms of a discrepancy between verbal and reasoning functions (Inglis, 1958); and Cattell (1963) would surely anticipate an age decrement in "fluid intelligence" (R-related) long

---

[4] The writer is indebted to Campbell for a timely suggestion along this line.

TABLE 8

ESTIMATED "TRUE" OR INTRINSIC CHANGES IN ALPHA SCORE, 1919 TO 1961

Subtest	Apparent gain 1919–1961	Cultural change ±	True gain[a]	
			Subtest	Component
Practical judgment	.42	.18*	.24	.37**(V)
Synonym-antonym	.50	.00	.50	
Disarranged sentences	.60	.88**	−.28	
Information	.97	−.03	1.00	
Arithmetical problems	−.34	−.31**	−.03	−.18*(N)
Number series completion	−.08	.63**	−.71	
Following directions	.05	.03	.02	−.61**(R)
Analogies	.16	.90**	−.74	
Total	.50	.44**		.06

[a] M of 101 male freshmen of 1961 minus M of 96 male freshmen of 1919.

* = $p < .05$.

** = $p < .01$.

before any decrement in "crystallized intelligence" (V-related).

A second matter requiring comment concerns possible interpretations of relationships displayed in Table 7 and not anticipated in the stated hypotheses. The first P_2 column contains one of these, a verbal loading of −.50 on the Egocentric Independence factor. One is tempted to speculate that high scorers on this factor, as defined by the items, would be less involved in social interactions, voluntarily and involuntarily, than others. If so, less breadth of general vocabulary could well be a consequence.

A second complex of relationships worth a comment appears in the same table under the Numerical Component and opposite Factors III and IV, Physical Vigor and Egocentric Independence. The question is *why* high scorers on these factors should tend to improve their N scores more than low scorers on the factors. A possible answer would seem to be because they manipulate the problem more vigorously. In the context of sex differences in problem-solving, Witkin et al., (1954) have noted the superiority of males at perceptual problems, and have observed that it seems to stem from a more aggressive, vigorous, and persistent attack. If it is true, within a sex group, that the possession of male characteristics argues for a more masculine and a superior performance, then it is pertinent to observe that physical vigor and egocentric independence are appropriately masculine characteristics.

Still with reference to Table 7, it seems both important and appropriate to note that the factors which argue for improvement in V over time are not, apparently, the same as those which argue for improvement in N or R. Indeed, considerable specificity would seem to be the rule.

Evaluation

Many of the problems attendant upon the conduct of an investigation such as the present one have been priorly dealt with (Owens, 1953) and the writer has nothing of substance to add. However, two which persist and recur should be treated here.

First, there is the ever-present question of a bias in the sampling of *Ss*. Of the 127 *Ss* of 1950, 96 were

TABLE 9

TOTAL AND COMPONENT ALPHA DIFFERENCE SCORES OF TESTED AND UNTESTED SUBJECTS (1950–1919)

Component	$\overline{D_T}$	$\overline{D_U}$	t^a
Verbal	.631	.679	.02
Numerical	−.059	−.151	.77
Reasoning	.150	−.035	1.23
Total	.583	.447	1.49

Note.—$M = 5$; $SD = 1$. For tested *S*s, $N = 96$; for untested *S*s, $N = 31$.

[a] None of these values is significant.

actually retested; 13 were deceased, 8 refused retesting, 5 were disabled by accident or illness, and 5 could not be located. Although much of it was unavoidable, the total loss was 31 cases, and the question of how this loss might have affected the reported shifts or changes in Alpha scores is a legitimate one. Accordingly, Table 9 contains what seems to be the most direct evidence available on this point. It shows the difference scores for 1950 minus 1919 of those who were tested and untested in 1961. The mean increment in total Alpha score actually obtained by the 127 *S*s of 1950 is .55σ. If this is compared with the mean obtained by the 96 *S*s retested in 1961, it will be noted that the discrepancy is of the order of .03σ. Thus, had the present bias existed at the initial retesting, this would have been its magnitude. This is clearly *not* an estimate of the bias introduced into the difference scores of 1961–1950. However, it should overestimate the later bias, since gains were much larger at the first retesting than at the second.

The second persisting problem is one of the extent to which the speed requirements of the Alpha test have conditioned the results. In anticipation of an attempt to evaluate this problem, all *S*s were given the examination on both standard time and double time bases. This was done in such fashion as to keep the two scores independent, and to avoid any alteration in the prescribed procedure for the initial and standard administration. The method outlined by Helmstadter and Ortmeyer (1953) was then employed to indicate whether or not a given subtest was significantly speeded, significantly powered, or neither. In essence, the method utilizes items omitted on standard time as an index to speeding, and items wrong on double time as an index to power. Given these two indices, one may subtract the latter from the former, *S* by *S*, and generate a single distribution of scores to be tested directly for the significance of the departure of its mean from that of a distribution with mean zero.

Table 10 contains a summary resulting from the application of this method to the data of both earlier and later retestings. It will be noted that there are both more speeded and more powered subtests in 1961 than in 1950. Of those which appear for the first time in 1961, Number Series

TABLE 10

SPEED AND POWER INDICES ON THE ARMY ALPHA SUBTESTS

Subtest	Speed-Power composition (1950)	Speed-Power composition (1961)
3. Practical judgment	S S[a]	S S
7. Analogies	S S	S S
4. Synonym-antonym	S	S S
6. Number series completion		S
5. Disarranged sentences		
2. Arithmetic problems		P
8. Information	P P[b]	P P

[a] S S = Speeded, *t* test significant at .01; S = Speeded, *t* test significant at .05.

[b] P P = Power, *t* test significant at .01; P = Power, *t* test significant at .05.

Completion is speeded, Arithmetical Problems is powered, and both load on the N component. In addition, the subtest on which the '50 to '61 apparent loss was largest, Arithmetical Problems, became more powered. It thus seems difficult to relate changes in Alpha test performance to changes in its speed-power composition.

In concluding, two incidental implications of this study seem noteworthy. First, cross-sectional and longitudinal studies of the effects of age on mental abilities have yielded divergent and apparently contradictory results to data. Table 8 illustrates a rationale for their useful and meaningful integration. Second, the demonstration that patterns of living moderate the relationship of age to mental ability implies that cognitive decline, like cognitive development, is conditioned to some extent on the nature and intensity of environmental stimulation.

REFERENCES

BAYLEY, N., & ODEN, M. The maintenance of intellectual ability in gifted adults. *Journal of Gerontology,* 1955, **10,** 91–107.

BENTZ, V. J. A test-retest experiment on the relationship between age and mental ability. *American Psychologist,* 1953, **8,** 319–320.

BIRREN, J. E. *The psychology of aging.* Englewood Cliffs, N. J.: Prentice-Hall, 1964.

BROZEK, J., & TIEDE, K. Reliable and questionable significance in a series of statistical tests. *Psychological Bulletin,* 1952, **49,** 339–341.

CAMPBELL, D. T. From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in measuring change.* Madison, Wisc.: The University of Wisconsin Press, 1963. Pp. 212–242.

CARROLL, J. B. Biquartimin criterion for rotation to oblique simple structure in factor analysis. *Science,* 1957, **126,** 1114–1115.

CATTELL, R. B. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology,* 1963, **54,** 1–22.

CHARLES, D. C. Ability and accomplishment of persons earlier judged mentally deficient. *Genetic Psychology Monographs,* 1953, **47,** 3–71.

GUILFORD, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

HEBB, D. O. *Organization of behavior.* New York: Wiley, 1949.

HELMSTADTER, G., & ORTMEYER, D. Some techniques for determining the relative magnitude of speed and power components of a test. *Educational and Psychological Measurement,* 1953, **13,** 280–287.

HUNT, J. McV. *Intelligence and experience.* New York: Ronald Press, 1961.

JONES, H. E. Intelligence and problem solving. In J. E. Birren (Ed.), *Handbook of aging and the individual.* Chicago: The University of Chicago Press, 1959. Pp. 700–735.

KALLMAN, F. J., & JARVIK, L. Individual differences in constitution and genetic background. In J. E. Birren (Ed.), *Handbook of aging and the individual.* Chicago: The University of Chicago Press, 1959. Pp. 249–253.

KUHLEN, R. G. Social change: A neglected factor in psychological studies of the life span. *School and Society,* 1940, **52,** 14–16.

MORGAN, W. A. A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika,* 1939, **31,** 13–19.

NISBET, J. D. Intelligence and age: Retesting with twenty-four years' interval. *British Journal of Educational Psychology,* 1957, **27,** 190–198.

OWENS, W. A. Age and mental abilities: A longitudinal study. *Genetic Psychology Monographs,* 1953, **48,** 3–54.

PITTMAN, E. G. A note on normal correlation. *Biometrika,* 1939, **31,** 9–12.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

WECHSLER, D. Cognitive, conative and non-intellective intelligence. *American Psychologist,* 1950, **5,** 78–83.

WITKIN, H. A., LEWIS, H. B., HERTZMAN, M., MACHOVER, K., MEISSNER, P. B., & WAPNER, S. *Personality through perception.* New York: Harper & Row, 1954.

(Received December 17, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 6, 326–336

STRUCTURE OF INTELLIGENCE IN NEGRO AND WHITE CHILDREN[1]

IRA J. SEMLER AND IRA ISCOE[2]
Cedar Rapids Community School District, Iowa
University of Texas

Performance levels on the WISC and on the Progressive Matrices (PM) and the factorial structures of the WISC were compared for white and Negro children. WISCs were administered individually to 141 white and 134 Negro Ss ranging from 5 through 9 yrs. of age. The 7-, 8-, and 9-yr.-old Ss were also given the PM as a group test. White Ss had higher WISC FS IQs than Negroes at all age levels ($p < .001$) and higher PM total scores at the 7-yr. level ($p < .05$) but not at the 8- and 9-yr. levels. While the white and Negro WISC intercorrelation matrices appeared to be similar, a multivariate analysis of variance showed heterogeneous dispersions ($p < .001$) and mean vectors ($p < .001$). Since the multivariate analysis indicated statistically dissimilar WISC structures for the white and Negro samples, separate factor analyses were computed to explore the nature of the differences. The PM intercorrelation matrices for whites and Negroes were highly similar.

For the past several years, research evidence has been increasing steadily on those psychological variables which are presumably indicators of performance in school and society. The relatively inferior performance of Negroes compared with whites on both individual and group intelligence tests, for example, has been documented extensively (Dreger & Miller, 1960; Garrett, 1947; Shuey, 1958; Young & Bright, 1954); however, the rationale for these differences continues to be a controversial topic in the literature. Although Garrett (1947) and Shuey (1958) have offered genetic explanations for the observed differences, the importance of the psychocultural environment and social learning in early childhood is being stressed by others (Dreger & Miller, 1960). It is highly probable that changes in economic conditions, early intervention with innovative educational procedures, and increased opportunity to learn will produce a different yield from cross-cultural studies of intelligence than has been reported previously. While comparative studies of intelligence are hazardous, seriatim measurements taken in conjunction with new educational procedures instituted for "inferior" children would have some usefulness in evaluating the results of changing psychocultural influences.

The choice of research instruments in comparative studies of intelligence is extremely important since the constructs and operations defining "intelligence" may vary markedly from test to test and may be related in different ways to the between-race effect under study. Also, the complicated relationships among intelligence, learning, and the ability to cope with environmental conditions are at best unclear. Further, functional intelligence or the intelligence needed to survive in a certain environment under certain conditions involves attributes not sampled by some tests. For example, Young and Bright (1954) have suggested that the Wechsler Intelligence Scale for Children (WISC) is not

---

[1] This research was partially supported by a grant from the Human Genetics Foundation.

[2] Special appreciation is expressed to Diane DeLancey for computational help and preparation of tables and manuscript.

suitable for Southern Negro children since they seem to be functioning effectively in their environment even though they make significantly lower IQs than white children. This suggests that the mental ability estimates provided by some instruments may be inappropriate for individuals who, for whatever reason, have need for a functional intelligence quite different from that needed by the standardization population. This is no argument for the so-called "culture fair" test of intelligence since it is almost impossible to construct and it certainly misses the point in intellectual assessment (Dreger & Miller, 1960); however, there is need for further comparative research designed to explore differences in measured intelligence as a function of test instrument and changing psychocultural conditions.

The principles underlying the construction of some tests come closer to the functional intelligence concept than others and may provide a much different assessment as a consequence. For example, the WISC supposedly taps general intelligence but may carry varying loads of memory, drive, and nonintellective factors (Wechsler, 1949). Littell's (1960, p. 144) review suggests that the WISC Verbal (V) and Performance (P) scales are not factorially pure but contain additional factors which might be labeled general eductive ability and memory. In particular, the WISC V scale contains factor loadings which reflect previous experience and education (Cohen, 1959). In contrast, the Progressive Matrices (PM) was based on Spearman's cognitive principles (Raven, 1947) and was designed to be less susceptible to the influence of previous learning and experience. Since the PM requires the ability to observe and perceive relationships within the boundaries of the testing situation itself, it may provide results more closely related to functional intelligence than the WISC.

The present study extends some previous research (Semler & Iscoe, 1963) which found that the paired-associate (PA) learning (pairs of objects and colored pictures of these same objects) of Southern white and Negro children did not differ significantly at the older age levels studied (8- and 9-year-olds) despite the significantly lower WISC Full Scale (FS) IQs of the Negroes. This study analyzes the performance of these same white and Negro *S*s on the PM at the 7-, 8-, and 9-year age levels for comparison with the analysis of the effect of race on WISC performance. Since the PM requires within-test learning involving cognitive processes and ostensibly does not carry a heavy psychocultural loading, it was hypothesized that between-race performance differences would be less on this test than on the WISC. Comparing the intercorrelation matrices of the white and Negro samples for the WISC and for the PM was of major importance in this study. The basic analyses tested the similarity of intercorrelations within each cell of the combined white-Negro matrices. Further, multivariate analysis of variance and factor analysis procedures were used to test the similarity of WISC structure for the white and Negro samples.

METHOD

Subjects

The subjects (*S*s) used in this investigation were 134 Negro and 141 white children ranging from 5 to 9 years of age who were enrolled in the public schools and kindergartens in Austin, Texas.[3] At the time of the study, these schools were still segregated as a result of race differences in housing location rather than school policy.

---

[3] Appreciation is expressed to the administrative officials of the Austin Public School System, Texas, for their cooperation in conducting this study.

Procedure

Each child had been administered the WISC by a trained examiner as part of a study of race differences in learning ability (Semler & Iscoe, 1963). The five WISC V subtests given were: Information (I), Arithmetic (A), Similarities (S), Vocabulary (Voc), and Digit Span (DS), to provide a WISC V IQ. The four WISC P subtests given were: Picture Completion (PC), Block Design (BD), Object Assembly (OA), and Coding (C), which were prorated to give a WISC P IQ. All 275 *S*s were tested for this portion of the study.

Partly as a result of finding that white children had higher WISC FS IQs compared with Negro children at all age levels in the previous study (Semler & Iscoe, 1963) and to provide comparative data on a less culturally loaded test, sets A, AB, and B of the Coloured PM (Raven, 1956) were administered as group (classroom) tests to the 160 *S*s who were 7, 8, and 9 years of age. The 5- and 6-year-olds were not administered the PM because of their inability to handle the test in a group setting.

Statistical Design

The statistical design was carried out in three phases. The first phase analyzed white-Negro differences in measured intelligence on the two instruments used. Separate race by age proportional analyses of variance (Lindquist, 1953) were carried out on WISC FS IQs and PM total scores providing comparative data within *S*s on the two instruments.

In the second phase, all possible zero order correlations within and between the WISC and PM subtests and totals were computed separately for the Negro and white samples without regard to age. The resulting white and Negro intercorrelation matrices were combined with the appropriate parts of the WISC standardization population intercorrelation matrix for the 7½-year-old level (Wechsler, 1949, p. 10, Table IV) for comparison. Each cell of the combined matrix contained a set of three intercorrelations which were tested for homogeneity (Edwards, 1960) to determine within-set intercorrelation similarity. This type of analysis also allowed separate tests of the homogeneity of pairs of intercorrelations within sets, that is, for white versus Negro, white versus WISC standardization population, and Negro versus WISC standardization population intercorrelation pairs. The same type of analysis was used to test the homogeneity of white-Negro intercorrelation sets among PM subtests A, AB, and B and the correlations between the WISC and the PM subtests and totals.

The third phase compared the structure of WISC intelligence for white and Negro children with a multivariate analysis of variance and with exploratory factor analyses.[4]

RESULTS

Race by Age Proportional Analysis of Variance of WISC FS IQs

The means and standard deviations of WISC FS IQs by sex for white and Negro children at five age levels are presented in Table 1. A race by age proportional analysis of variance was computed after randomly discarding *S*s to equalize cell numbers ($N = 24$). The highly significant race difference obtained in favor of white children ($F = 64.89$, $df = 1/230$, $p < .001$) was supported at each of the age levels by an analysis of the simple effects using the mean square within-subgroups as the error term. The age effect was significant ($F = 6.98$, $df = 4/230$, $p < .001$), indicating different mean WISC FS IQs for the combined racial groups at the various age levels analyzed. The significant Race × Age

---

[4] The help and advice of H. P. Bechtoldt, University of Iowa, in carrying out this phase of the analysis is gratefully acknowledged. Bechtoldt's modification and extension of the multivariate analysis of variance (MULTAN) program of Cooley and Lohnes (1962) was used to test the hypotheses of constant variances and constant covariances, homogeneity of dispersion, and homogeneity of mean vectors. Further, Bechtoldt's comprehensive factor-analysis (FACTAN) program was used to explore the structure of WISC intelligence for white and Negro children separately. The MULTAN and FACTAN programs are described completely in the User's Guide, University Computer Center, University of Iowa.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF WISC FS IQs FOR WHITE (N = 141) AND NEGRO (N = 134) CHILDREN AT FIVE AGE LEVELS

*S*s	Age in years					Total
	5	6	7	8	9	
White						
Boys						
N	19	22	12	15	12	80
M	110.5	104.2	97.8	105.1	98.6	
σ	11.5	15.0	12.2	15.6	13.9	
Girls						
N	7	15	15	12	12	61
M	116.6	104.9	97.3	105.8	94.3	
σ	7.9	6.3	7.7	15.2	16.4	
Negro						
Boys						
N	13	14	12	20	12	71
M	83.5	92.6	90.7	98.3	89.6	
σ	13.1	11.3	8.1	9.9+	7.8	
Girls						
N	13	12	12	14	12	63
M	82.5	103.2	90.8	96.6	85.1	
σ	13.9+	7.2	12.3	11.2	11.3	

interaction ($F = 8.79$, $df = 4/230$, $p < .001$) indicated that the race differences in WISC FS IQs found in the simple effects analysis were not the same at each of the age levels studied.

WISC Intercorrelations: Comparison of White and Negro Samples and WISC Standardization Population

All possible linear correlations among WISC subtest scaled scores and V IQ, P IQ, and FS IQ are reported for the white and Negro samples and the WISC standardization population in Table 2 in combined matrix form. Each cell in the combined matrix contains a set of three intercorrelations, one for each of the three comparison groups. The set of intercorrelations in each cell was tested for homogeneity and z transformations were used to test the statistical difference between all possible pairs of intercorrelations within each of the heterogeneous sets. Only one set of WISC V subtest intercorrelations, A versus Voc, was significantly heterogeneous ($\chi^2 = 10.33$, $df = 2$, $p < .01$) with a significantly higher ($z = 3.03$, $p < .005$) relationship for white *S*s ($r = .52$) than Negroes ($r = .21$). A most striking finding was that all WISC P subtest intercorrelation sets were homogeneous even though slight differences between correlations within sets are evident. That part of the WISC structure defined by the intercorrelations between V and P subtests showed that five of the 20 intercorrelation sets exhibited significant heterogeneity. Three of the heterogeneous sets involved A with BD ($\chi^2 = 9.47$, $df = 2$, $p < .01$), A with OA ($\chi^2 = 8.24$, $df = 2$, $p < .01$), and A with C ($\chi^2 = 12.86$, $df = 2$, $p < .01$) with significantly higher intercorrelations for white *S*s than Negroes. The S versus PC intercorrelation set was heterogeneous ($\chi^2 = 6.65$, $df = 2$, $p < .05$) due to the higher ($z = 2.57$, $p < .01$) correlation for Negroes ($r = .51$) compared to the WISC standardization population ($r = .27$).

Of further interest are the part-whole correlation comparisons between WISC V subtests and V IQ for the three groups. In three sets (V IQ versus S, Voc, and DS) there were no white-Negro differences and heterogeneity was due to the significantly lower correlations of the WISC standardization population. In another set, V IQ versus A, the white correlation ($r = .73$) contributed significantly ($\chi^2 = 9.13$, $df = 2$, $p < .05$) to set heterogeneity with no difference between the Negro ($r = .57$) and WISC standardization population ($r = .55$) correlations. All of the part-whole correlation sets between WISC P subtests and P IQ were heterogeneous due to the significantly higher white and Negro correlations. In the P IQ versus A set, heterogeneity ($\chi^2 = 12.36$, $df = 2$, $p < .01$) was due to the significantly lower intercorrelation of

TABLE 2

INTERCORRELATIONS AMONG WISC SUBTEST SCALED SCORES, V IQ, P IQ, AND FS IQ FOR WHITE (N = 141) AND NEGRO (N = 134) CHILDREN AND WISC STANDARDIZATION POPULATION

WISC Subtests	*Ss*	I	A	S	Voc	DS	PC	BD	OA	C	V IQ	P IQ	FS IQ
Information (I)	White		.44	.48	.59	.31	.32	.35	.23	.17	.77	.38	.64
	Negro		.42	.47	.42	.28	.44	.22	.27	.32	.71	.45	.66
	WISC		.51	.49	.55	.39	.24	.33	.27	.26	.64	.44	.59
Arithmetic (A)	White			.36	.52*	.42	.30	.53*	.41**	.40*	.73**	.56*	.72*
	Negro			.32	.21	.23	.23	.28	.09	.01	.57	.21	.45
	WISC			.40	.46	.40	.29	.27	.29	.32	.55	.46	.57
Similarities (S)	White				.49	.22	.36**	.28	.21	.12	.72*	.35	.59
	Negro				.55	.32	.51	.18	.11	.20	.77	.35	.64
	WISC				.45	.34	.27	.29	.29	.15	.55	.41	.53
Vocabulary (Voc)	White					.38	.41	.35	.42**	.26	.82*	.52	.74
	Negro					.36	.43	.24	.15	.28	.74	.39	.65
	WISC					.38	.36	.33	.30	.22	.66	.47	.63
Digit Span Total (DS)	White						.10	.42	.36	.27	.63**	.39	.57*
	Negro						.27	.23	.23	.12	.60	.30	.51
	WISC						.23	.29	.13	.24	.44	.29	.42
Picture Completion (PC)	White							.32	.37	.18	.41	.71*	.64*
	Negro							.29	.43	.20	.53	.70	.71
	WISC							.28	.28	.12	.42	.34	.43
Block Design (BD)	White								.53	.40	.51	.75*	.71**
	Negro								.40	.21	.30	.63	.53
	WISC								.53	.26	.42	.53	.52
Object Assembly (OA)	White									.26	.44	.78*	.69
	Negro									.32	.28	.80	.62
	WISC									.30	.38	.59	.52
Coding (C)	White										.32	.59*	.52
	Negro										.27	.58	.49
	WISC										.31	.32	.35
Verbal IQ (V IQ)	White											.59	.89
	Negro											.52	.87
	WISC											.60	.90
Performance IQ (P IQ)	White												.89
	Negro												.87
	WISC												.89
Full Scale IQ (FS IQ)	White												
	Negro												
	WISC												

* $p < .01$ that the set of three r's tabled in the cell are estimates of the same population.

** $p < .05$ that the set of three r's tabled in the cell are estimates of the same population.

Negroes ($r = .21$) compared with whites ($r = .56$).

WISC Structure: Multivariate Analysis of Variance and Exploratory Factor Analyses

It should be noted that the WISC was not designed for factor analysis; however, an exploratory study was made to provide comparative data on the WISC structure for the white and Negro samples. Prior to the factor analysis, a multivariate analysis of variance was computed to evaluate the hypothesis that these data might represent two independent random samples from a single multivariate normal population with reference to the two sets of nine means and the two variance-covariance matrices. If this hypothesis is tenable at least with respect to the variance-covariance matrices (or dispersion matrices), a single factor analysis based on the average within-groups correlation matrix would be indicated (Cooley & Lohnes, 1962). The hypothesis of homogeneous dispersions was not tenable ($F = 1.99$, $df = 45/24537.436$, $p < .001$). Further, the hypothesis of the homogeneity of mean vectors (equality of population centroids) was not supported ($F = 14.99$, $df = 9/265$, $p < .001$). Sufficient statistical evidence existed at this point to indicate that dissimilar WISC factor structures would be expected for the white and Negro samples.

Since a single factor analysis based on the average within-group correlation matrix is contraindicated, two separate exploratory factor analyses were run using the maximum likelihood and Varimax rotation (Kaiser, 1959) procedures. The hypothesis of independence was refuted for each set of data; the tests for this hypothesis were significant for the white data ($\chi^2 = 385.91$, $df = 36$, $p < .001$) *and* for the Negro data ($\chi^2 = 291.74$, $df = 36$, $p < .001$). It follows from these statistical tests that a separate factor analysis for each set of data could be justified. The maximum likelihood procedure applied to the data from the 141 white *S*s indicated only three significant factors from the test of the third factor residuals ($\chi^2 = 12.63$, $df = 12$). The data for the 134 Negro *S*s, however, showed five significant factors which appeared to be a function of the program procedure in which the early factor values were not stabilized. In a rerun with the program limited to three factors, the results indeed stabilized with the second factor residuals just short of significance at the .05 level. The third run, reported here, was then made in order to produce results more comparable to those of the white sample. The third factor residuals were clearly not significant ($\chi^2 = 12.22$, $df = 12$). The factor matrices for whites and Negroes are shown in Table 3. The comparability of factoring is indicated in part by the nearly equal properties of the total variance; the sums of the communalities over nine tests were 4.648 and 4.657 as shown on the bottom line of Table 3 and the proportion of total variance accounted for was 51.64% and 51.74% for the white and Negro samples, respectively. Only the Varimax rotation, as shown in Table 4, was made for the white and Negro samples, although an inspection of the graphical plots of each set of data indicated that an oblique structure would provide a better fit. However, in view of the inappropriateness of the nine WISC subtests for a factor analysis and the forcing procedure used to factor the Negro data, further transformations could hardly be justified.

Rationalizing the extracted factors by inferring sets of psychological abilities from inspection of the WISC

TABLE 3
MAXIMUM LIKELIHOOD FACTOR MATRICES FOR WHITE (N = 141) AND NEGRO (N = 134) CHILDREN

WISC Subtests	White				Negro			
	Factors				Factors			
	I	II	III	h²	I	II	III	h²
Information	.639	.331	−.167	.545	.641	.139	−.085	.437
Arithmetic	.736	−.137	−.175	.590	.465	.381	−.797	.996
Similarities	.536	.340	−.101	.413	.690	.340	.167	.620
Vocabulary	.748	.313	−.004	.658	.643	.233	.228	.519
Digit Span Total	.535	−.157	−.052	.313	.458	.046	−.002	.211
Picture Completion	.483	.162	.120	.274	.694	−.054	.086	.492
Block Design	.669	−.333	.001	.558	.429	−.162	−.178	.242
Object Assembly	.673	−.244	.687	.984	.572	−.778	−.145	.953
Coding	.440	−.307	−.155	.312	.378	−.156	.133	.185
Σa^2	3.408	.658	.582	4.648	2.862	.995	.800	4.657

subtests exhibiting nonzero loadings has little meaning in this study since the white and Negro structures were significantly different, although some general comparisons may be made. The WISC I, S, Voc, and PC subtests load Factor I for both the white and Negro samples, suggesting similar clusters of verbal comprehension abilities. While the WISC OA subtest loaded Factor III for the white sample and Factor II for the Negro sample, reflecting the ability to organize visually perceived materials, the Negro factor was quite complex due to the number of indeterminant factor loadings. Only the WISC A subtest loaded heavily on Factor III for the Negro sample, leaving a "clean" factor representing that aspect of verbally retained knowledge gained only from formal education. Factor II for the white sample was a complex factor involving loadings of the WISC A, Voc, PC, and C subtests.

TABLE 4
VARIMAX ROTATION FACTOR MATRICES FOR WHITE (N = 141) AND NEGRO (N = 134) CHILDREN

WISC Subtests	White				Negro			
	Factors				Factors			
	I	II	III	h²	I	II	III	h²
Information	.696	.244	.034	.545	.536	.227	.314	.437
Arithmetic	.426	.624	.141	.590	.194	.033	.978	.996
Similarities	.623	.148	.054	.413	.767	.048	.171	.620
Vocabulary	.733	.263	.227	.658	.710	.105	.066	.519
Digit Span Total	.257	.463	.180	.313	.385	.199	.154	.211
Picture Completion	.426	.149	.265	.274	.570	.394	.112	.492
Block Design	.213	.646	.308	.558	.199	.385	.233	.242
Object Assembly	.177	.326	.920	.984	.068	.973	.041	.953
Coding	.100	.544	.075	.312	.294	.309	−.058	.185

Race by Age Proportional Analysis of Variance of PM Total Scores

Table 5 presents the means and standard deviations of PM total scores (sum of scores on sets A, AB, and B) by sex for white and Negro children at the 7-, 8-, and 9-year age levels. As with the WISC, a proportional analysis of variance was carried out after randomly discarding *Ss* to equalize cell numbers (N = 24). The main effects of race ($F = 4.44$, $df = 1/138$, $p < .05$) and age ($F = 20.64$, $df = 2/38$, $p < .001$) were significant. Analysis of the simple effects for age indicated a significant difference in PM total scores favoring white children at the 7-year-old level ($t = 1.99$, $df = 50$, $p < .05$) but *not* at the 8- and 9-year-old levels. While the 7-year-old

TABLE 5

MEANS AND STANDARD DEVIATIONS OF PM TOTAL SCORES FOR WHITE ($N = 78$) AND NEGRO ($N = 82$) CHILDREN BY SEX AT THREE AGE LEVELS

Ss	Age in years			Total
	7	8	9	
White				
Boys				
N	12	15	12	39
M	15.83	25.20	21.42	
σ	8.52	8.18	5.88	
Girls				
N	15	12	12	39
M	15.00	22.33	22.08	
σ	8.16	5.34	5.14	
Negro				
Boys				
N	12	20	12	44
M	9.42	21.85	21.75	
σ	9.53	5.19	5.86	
Girls				
N	12	14	12	38
M	12.67	20.57	19.50	
σ	9.89	4.99+	5.69	

Ss contributed to the significant race differences reported above for PM total scores, the Race × Age interaction was not significant.

PM Intercorrelations: Comparison of White and Negro Samples

The intercorrelations among scores on PM sets A, AB, B, and total for white and Negro children are reported in Table 6. All intercorrelation sets were homogeneous. However, one part-whole correlation set, PM AB versus PM total, was significantly heterogeneous ($z = 3.39$, $p < .001$) with a higher correlation for whites ($r = .91$) than for Negroes ($r = .80$).

Interrelatedness of WISC and PM: Comparison of White and Negro Samples

Table 7 presents the correlations between the WISC subtests and IQs with PM sets A, AB, B, and total for white and Negro *S*s. In general, correlations were small with set heterogeneity evidenced for PM AB versus WISC I, A, and V IQ where the correlations were significantly higher for white *S*s than for Negroes. The correlations between WISC FS IQ and PM total scores for whites ($r = .47$) and Negroes ($r = .34$) were not surprising in view of the many reasons for a moderate or even low relationship. It should be noted, however, that previous studies (Barratt, 1956; Martin & Weichers, 1954; Stacey & Carleton, 1955) have reported generally higher correlations between the WISC and the PM.

DISCUSSION

One of the guiding hypotheses in this study was that white-Negro differences in level of measured intelligence would be less on a test involving cognitive abilities than on a test loaded with psychocultural factors and this was partially borne out. Significantly lower WISC FS IQs were obtained for the Negro children compared with

TABLE 6

INTERCORRELATIONS AMONG PM SETS A, AB, B AND TOTAL SCORES FOR WHITE ($N = 78$) AND NEGRO ($N = 82$) CHILDREN

PM Sets	Ss	A	AB	B	Total
A	White		.64	.54	.80
	Negro		.49	.46	.75
AB	White			.65	.91*
	Negro			.57	.80
B	White				.86
	Negro				.86
Total	White				
	Negro				

* $p < .01$ that the set of two r's tabled in the cell are estimates of the same population.

TABLE 7

CORRELATIONS BETWEEN WISC SUBTESTS AND IQS AND PM SUBTESTS AND TOTAL SCORES FOR WHITE ($N = 78$) AND NEGRO ($N = 82$) CHILDREN

WISC	*S*s	PM			
		A	AB	B	Total
Information	White	.27	.26**	.20	.28
	Negro	.07	.01	.07	.10
Arithmetic	White	.20	.29**	.30	.31
	Negro	.12	.04	.14	.11
Similarities	White	.29	.40	.38	.42
	Negro	.27	.20	.27	.31
Vocabulary	White	.29	.34	.19	.31
	Negro	.21	.23	.28	.29
Digit Span Total	White	.17	.19	.15	.20
	Negro	.07	.08	.10	.14
Picture Completion	White	.23	.30	.26	.31
	Negro	.20	.23	.14	.23
Block Design	White	.37	.41	.33	.43
	Negro	.22	.28	.30	.33
Object Assembly	White	.26	.30	.26	.32
	Negro	.12	.26	.18	.23
Coding	White	.09	.14	.16	.15
	Negro	−.08	.00	.01	−.002
Verbal IQ	White	.33	.40*	.33	.42
	Negro	.25	.19	.27	.31
Performance IQ (prorated)	White	.33	.40	.35	.42
	Negro	.18	.29	.23	.30
Full Scale IQ	White	.37	.45	.38	.47
	Negro	.24	.27	.29	.34

* $p < .01$ that the set of two *r*'s tabled in the cell are estimates of the same population.

** $p < .05$ that the set of two *r*'s tabled in the cell are estimates of the same population.

white children at each of the age levels studied; however, no significant differences were obtained on PM total scores for these same Negro and white children at the 8- and 9-year age levels. Significantly lower PM total scores were found for Negroes at the 7-year-old level suggesting that age and/or experiential factors related to age may be influential in cognitive development. The findings may be explained partially by the differential demands of the two instruments. WISC performance is influenced by verbal comprehension and experience with verbal manipulation, the ability to tolerate distraction including time pressures, and previous experience with materials designed to develop manipulative skills and perceptual organization. In contrast to the Negro children, these white *S*s had ample opportunity for the early development of verbal skills and the acquisition of a fund of social information as well as early life experiences with a variety of play materials, all of which contribute to the development of those mental abilities sampled by the WISC. On the other hand, the PM was designed to tap those cognitive abilities which are less susceptible to individual differences in prior experience and learning and the environmental background of the Negroes did not operate to reduce performance on this test as was the case on the WISC.

In view of the finding that the performance of the older Negro and white *S*s did not differ on the PM, it might be useful to point to previous research findings (Semler & Iscoe, 1963) showing that the PA learning of these same 8- and 9-year-old Negro and white children did not differ despite significantly lower WISC FS IQs for the Negroes. The PA materials were objects commonly found in the environment of most children regardless of socioeconomic condition. The PA task required *S*s to associate pairs of objects or colored pictures of these objects by the anticipation method. The object and picture pairs were arranged in conceptually similar and dissimilar sets and required *S* to learn new associative responses. The PM also involves new learning within the testing situation by requiring *S* to observe the relationship of patterns within a perceptual field and make a correct association. This interpretation depends, of course, on the general similarity of the abilities required in the two situations but is offered as speculation for future research on the relationship between learning and measured intelligence.

The combined white-Negro intercorrelation matrices for the WISC and for the PM provided the basic data for the structure-of-intelligence comparisons. The white-Negro PM subtest intercorrelation matrices were quite similar, suggesting that the interrelationships among the stages of cognitive development sampled by this test are quite similar for the two races; however, this assumption was not subjected to further analyses due to the small number of subtests. Further research comparing the development of cognitive abilities for groups with varying psychocultural backgrounds seems promising.

The white-Negro WISC intercorrelation matrices appeared to be similar with the exception of the lower intercorrelations involving the A subtest for the Negroes. Performance on this subtest is directly related to verbally retained knowledge learned only through formal education. The multivariate analysis of variance indicated that statistically dissimilar WISC structures existed for the whites and for the Negroes and separate exploratory factor analyses were carried out. It should be noted that the WISC A subtest was practically a unitary factor for the Negroes but not for the whites. Interestingly enough, the cluster of verbal comprehension abilities was similar for both races. Unique factor-analytic procedures have been used for the WISC with both white and Negro children (Osborne, 1963, 1964); however, special statistical problems arise when group comparisons are made and the WISC is not designed for this purpose. It seems reasonable to suggest that future comparative factor-analytic research could well explore separately those domains of intellectual abilities theorized to be influenced by psychocultural variables and formal training.

REFERENCES

BARRATT, E. S. The relationship of the Progressive Matrices (1938) and the Columbia Mental Maturity Scale to the WISC. *Journal of Consulting Psychology*, 1956, **20**, 294–296.

COHEN, J. The factorial structure of the WISC at ages 7–6, 10–6, and 13–6. *Journal of Consulting Psychology*, 1959, **23**, 285–299.

COOLEY, W. W., & LOHNES, D. R. *Multivariate procedures for the behavioral sciences*. New York: Wiley, 1962.

DREGER, R. M., & MILLER, K. S. Comparative psychological studies of Negroes and whites in the United States. *Psychological Bulletin*, 1960, **57**, 361–402.

EDWARDS, A. L. *Experimental design in psychological research*. New York: Holt, Rinehart & Winston, 1960.

GARRETT, H. E. Negro-white differences in mental ability in the United States. *Scientific Monthly*, 1947, **65**, 329–333.

KAISER, H. F. Computer program for varimax rotation in factor analysis. *Educational and Psychological Measurements*, 1959, **19**, 413–420.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

LITTELL, W. M. The Wechsler Intelligence Scale for Children: Review of a decade of research. *Psychological Bulletin*, 1960, **57**, 132–156.

MARTIN, A. W., & WIECHERS, J. E. Raven's Colored Progressive Matrices and the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology*, 1954, **18**, 143–144.

OSBORNE, R. T. Factorial composition of the Wechsler Intelligence Scale for Children at the pre-school level. *Psychological Reports*, 1963, **13**, 443–448.

OSBORNE, R. T. WISC factor structure for normal Negro pre-school children. *Psychological Reports*, 1964, **15**, 543–548.

RAVEN, J. C. *Progressive Matrices (1947), Series A, AB, B*. London: H. K. Lewis & George G. Harrap, 1947.

RAVEN, J. C. *The Coloured Progressive Matrices, Sets A, AB, B (Revised Order 1956)*. London: H. K. Lewis, 1956.

SEMLER, I. J., & ISCOE, I. Comparative and developmental study of the learning abilities of Negro and white children under four conditions. *Journal of Educational Psychology*, 1963, 54, 38–44.

SHUEY, A. M. *The testing of Negro intelligence*. Lynchburg: J. P. Bell, 1958.

STACEY, C. L., & CARLETON, F. O. The relationship between Raven's Colored Progressive Matrices and two tests of general intelligence. *Journal of Clinical Psychology*, 1955, 11, 84–85.

WECHSLER, D. *Wechsler Intelligence Scale for Children (Manual)*. New York: Psychological Corporation, 1949.

YOUNG, F. M., & BRIGHT, H. A. Results of testing 81 Negro rural juveniles with the Wechsler Intelligence Scale for Children. *Journal of Social Psychology*, 1954, 39, 219–226.

(Received October 6, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 6, 337–340

DISCRIMINABILITY OF WORDS, AND LETTER CUES USED IN LEARNING TO READ[1]

S. J. SAMUELS[2] AND W. E. JEFFREY
University of California, Los Angeles

The 36 kindergarten Ss were distributed among 3 groups. Using the paired-associate anticipation method, each group was taught 1 of 3 lists of words that differed in discriminability. Discriminability was determined by the number of different letters (either 4, 6, or 8) used to construct 4 2-letter words. It was hypothesized that the more discriminable the list, the faster the learning rate but the greater the probability that Ss would learn on the basis of single letters. The hypothesis concerning learning rate was confirmed with $p < .01$. After the 4- and 8-letter groups were brought to the same criterion, the 8-letter group was found to identify words on the basis of single letter cues significantly more often ($p < .02$) than the 4-letter group, thus confirming the 2nd hypothesis as well.

Similarity among stimuli has not been shown to retard paired-associate learning (Underwood 1953a, 1953b) as much as one might expect from notions of the effect of stimulus generalization on learning (Gibson, 1940). This may be due to the fact that paired-associate learning is typically done with adult subjects (*S*s), and even though low-meaningfulness nonsense syllables may be used (Underwood, 1953a), the stimulus units (letters) are already well discriminated. If this were true, high-stimulus similarity in paired-associate learning with children should produce greater difficulty in that children would be less familiar with the letters, and thus increased similarity might produce results similar to those obtained by Gibson (1941) using nonsense figures with adult *S*s.

Most early reading texts select words on the basis of their frequency of usage. This increases the likelihood that the child is familiar with the sound of the word, and that the sound is meaningful. Also, as Silberman (1964) points out, the high-frequency words tend to be irregular and to differ markedly from each other in configuration and thus to be quite dissimilar. Furthermore, it is not unlikely that high similarity is purposely avoided as a potential source of confusion for the child. Although such procedures for word selection in teaching reading may facilitate initial learning, effective reading training should also provide the ability to transfer from the initial training such that the reader may emit correct sounds to new and unfamiliar combinations of letters. The extent to which this happens depends on what becomes the stimulus in the initial paired-associate learning situation. Anderson and Dearborn (1952, p. 215) suggest that "what children mainly seem to have an eye for is some detail of the word, a single letter or two which comes to stand for the whole word. . . ." Marchbanks and Levin (1965) provide support for this proposition in that they find that letters were used more often than word shape as a cue for recognition. If a single letter may serve as a cue for the whole word response, then rapid initial learning may indeed be a false economy. When a child provides a response to a new configuration of letters on the basis of a single letter cue, he may not even recognize that he has made an error until it becomes obvious from reading the rest of the sentence.

---

[1] This research was supported by United States Public Health Service Grant 07937 to the second author.

[2] Now at the University of Minnesota.

In a study that compared the relative transfer effects of word and letter training (Jeffrey & Samuels, in press) it was necessary to use common letters in each of two words in order to assure that the children utilized each specific letter. The results suggested that four words constructed of four letters were harder to learn than four words constructed of six letters, that is, learning was retarded by similarity. The present study was designed to test the relationship between the rate of learning and the number of letters, and, in addition, to evaluate the degree to which transfer is related to similarity as defined by number of letters used in the list. More specifically, we are testing the hypothesis that when a list is made up with no letters in common among the words, the children are more likely to identify new words on the basis of a single letter than when they are forced to utilize each letter in each word of the original list. To the extent that new words are identified on the basis of a single letter, transfer would of course be very poor.

METHOD

Subjects

The *Ss* were 36 kindergarten children from a Los Angeles elementary school and 24 children from a Santa Monica nursery school. The median age of both groups was 5 years 5 months.

Material for Learning Reading Responses

An artificial alphabet was created that had as little resemblance to English letters as possible. The responses to be associated with the letters were common English words assumed to be part of the *Ss'* repertoire (Figure 1). Three lists of two-letter words were constructed. List 4-L had four two-letter words constructed from only four different letters. List 6-L had four two-letter words constructed from six different letters. List 8-L had four two-letter words constructed from eight different letters. The words were printed on 5 × 8-inch index cards. These were presented serially by hand with an approximate 4-second response interval. Each word list was presented in three orders.

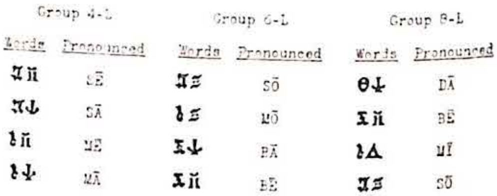


FIG. 1. Symbols used in each of three different lists and their pronunciations.

Material for Test of Cues Used in Learning Reading Responses

On each test card one of the artificial letters which *S* had seen in his previous training was now replaced with a letter that he had never seen before. In other words, the words were presented again, but this time with a new letter substituted for one of the letters in each word. This means Group 4-L was presented with 4 words, and Group 8-L was presented with 8 words. The new letter appeared half of the time in the first position and half of the time in the second position so that each original appeared once in combination with the new letter and the original letter always appeared in its original position. Thus, if *S* had learned to respond to any of the words in the original list on the basis of only one letter, he should not notice a change in at least half of these new pairs and he should provide the response he had previously learned to that letter without hesitation.

Procedure for Evaluating Effect of Number of Words on Acquisition

Twelve *Ss* were randomly assigned to each of the three different training lists. Each *S* was brought into the experimental room and told that he was going to play a game. After he was seated, he was told to select a small plastic toy, to put the toy in his desk, and that he could take the toy back to his room when he was finished playing the game. He was then shown the first four words in the list, told the word to be associated with the letters, and asked to repeat the word after each presentation. He was then told that for the rest of the trials he must try to say the correct word before the experimenter said the word.

All *S*s in the learning group were given 20 repetitions of the list and each correct response was recorded.

Procedure for the Transfer Test of Cues Used in Learning Reading Responses

Instructions were the same as for the learning groups. Twelve *S*s were assigned to each of two groups for training. One group was trained on Set 4-L, and the other on Set 8-L. Both groups were trained to a criterion of two successive correct trials. For the first 18 trials, the pairs were presented in a constant serial order. On the 19th and all subsequent trials, the order in which the word pairs were presented was randomized. This procedure facilitated learning markedly, particularly with the four-letter list. Upon reaching criterion, each *S* was shown the test cards serially, and asked, "Would you like to tell me what word this is?" His response was recorded and the next card presented.

RESULTS

For the acquisition groups the mean number of correct reading responses in 20 trials was 24.00 for Group 4-L, 38.25 for Group 6-L, and 54.00 for Group 8-L. Analysis of variance indicates a significant difference ($F = 19.05$, $df = 33$, $p < .01$). Specific comparisons between the means of Groups 4-L and 6-L, 4-L and 8-L, and 6-L and 8-L, using the Newman-Keuls procedure (Winer, 1962, pp. 80–85), indicated all of these differences to be significant ($p < .01$).

The two groups used for the transfer test were brought to criterion before the test. The technique of first presenting the pairs in a constant serial order for 18 trials proved to be very efficient. On Trial 19, Group 4-L had learned a mean of 2.9 pairs and Group 8-L had learned a mean of 3.2 pairs. Following the 18-trial sequence, the pairs were randomized. Group 4-L took a mean of 23.33 trials to criterion and Group 8-L a mean of 21.16 trials including the initial 18 trials of constant serial-order training. The performance on the transfer test was as predicted. Group 8-L made a significantly larger number of false identifications than did Group 4-L ($t_{22} = 2.60$, $p < .02$). Ten of 12 *S*s in Group 8-L made one or more false identifications on the transfer test as compared with only four of 12 *S*s in Group 4-L.

DISCUSSION

The hypothesis that stimulus similarity would prove to be a relatively important factor in paired-associate learning, if the stimuli are unfamiliar and children are used as *S*s, received strong support although it was also shown that with specialized training (constant serial-order), the effect of stimulus similarity on learning was reduced considerably. The results provide clear evidence that the number of *S*s who make identifications on the basis of a single letter increases with the number of letters on which they were trained, that is, increases with dissimilarity of stimuli. Thus, as hypothesized, training that forces attention to each letter is less likely to lead to subsequent reading errors than training which permits the child to identify words on the basis of a single feature. In an earlier study, Jeffrey and Samuels (1965) found that when children are given letter training first, *S*s had much less difficulty identifying the words in List 4-L than *S*s which were given word training with the words in List 6-L prior to learning List 4-L. In combination, then, these studies suggest that, in teaching children to read, it is important that the child learns to attend to all of the letters. This can be accomplished by either specific letter training or it may be accomplished by giving them experience with highly similar words which forces identification on the basis of more than a single stimulus feature.

Before accepting this generalization uncritically, it should be noted that the training situation used a highly restricted and unusual alphabet, and each child was trained individually by the experimenter. Furthermore, a small tangible incentive was provided to assure that *S* would be cooperative, and would remain in the experimental situation.

REFERENCES

ANDERSON, I. H., & DEARBORN, W. F. *The psychology of teaching reading*. New York: Ronald Press, 1952.

GIBSON, E. J. A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review*, 1940, 47, 196–229.

GIBSON, E. J. Retroactive inhibition as a function of degree of generalization between tasks. *Journal of Experimental Psychology*, 1941, 28, 93–115.

JEFFREY, W. E., & SAMUELS, S. J. The effect of method of reading training on initial learning and transfer. *Journal of Verbal Learning and Verbal Behavior*, in press.

MARCHBANKS, G., & LEVIN, H. Cues by which children recognize words. *Journal of Educational Psychology*, 1965, 56, 57–61.

SILBERMAN, H. F. *Exploratory research on a beginning reading program*. (Technical Memorandum No. TM-895/100/00) Santa Monica, Calif.: System Development Corporation, 1964, No. 29.

UNDERWOOD, B. J. Studies of distributed practice: VIII. Learning and retention of paired nonsense syllables as a function of intralist similarity. *Journal of Experimental Psychology*, 1953, 45, 133–142. (a)

UNDERWOOD, B. J. Studies of distributed practice: IX. Learning and retention of paired adjectives as a function of intralist similarity. *Journal of Experimental Psychology*, 1953, 45, 143–149. (b)

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received October 12, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 6, 341–346

SIX STRUCTURE-OF-INTELLECT HYPOTHESES IN SIX-YEAR-OLD CHILDREN[1]

R. E. ORPET AND C. E. MEYERS
California State College at Long Beach *University of Southern California*

A 20-test battery hypothesizing 6 Guilford structure-of-intellect cells was administered to 100 6-yr.-old kindergarten children. Principal component factors were given several varimax rotations. Hypotheses well sustained were memory for symbolic units (MSU), convergent semantic production (NM), divergent production of semantic units (DMU), and evaluation of figural units (EFU). A tentative factor was memory for symbolic units (MFU). Tests for an evaluation of symbolic units (ESU) factor loaded on EFU, suggesting that reaction to visual language symbols is still perceptual at age 6. Separation of MSU from MFU was believed due to visual vs. auditory channels.

The study of ability factors in young children has passed the stage of demonstrating that separate factors exist, with effort now devoted to systematic identification of measurable abilities. The present study is an extension of previous explorations made with 6-year-old preliterate subjects (*Ss*) (Loeffler, 1963; McCartin & Meyers, 1966; Meyers, Dingman, Orpet, Sitkei, & Watts, 1964; Meyers, Orpet, Attwell, & Dingman, 1962). Those studies had demonstrated separate primary abilities in motor, semantic, and figural domains, and in cognitive, convergent, divergent, and evaluative processes. The present study was designed to confirm some of the tentative factors and to extend the exploration into memory processes and symbolic content.

The structure-of-intellect (SI) language in this report reflects the model of human ability given by Guilford (1959) and by Guilford and Hoepfner (1963). In the SI, "cognitive" refers specifically to knowing or recognizing, not to all processes. It is distinguished from "convergent production," in which problem-solving is required, as in completing a verbal analogy. "Semantic" and "symbolic" differ in that the former deals in meanings of language systems while the latter deals in the symbols apart from their meaning, as in anagrams, algebra, or grammar.

HYPOTHESES AND TEST BATTERY

The hypothesized factors were developed primarily from Guilford's (1959) theoretical model of the structure of intellect and the child-development literature regarding the development and individuation of abilities. All of the hypothesized abilities have been well established at the adult level, but have not all been established by replicated factor studies as existing in childhood. The following criteria were used in the selection or development of instruments to measure the hypothesized abilities: (*a*) suitability for individual administration with 6-year-old children, (*b*) face validity of the test as a univocal measure of the hypothesized ability, (*c*) previous utilization in a factor study, and (*d*) efficiency in test time. Of the 20 tests used in the battery, some were used as previously standardized, while others were adapted for use with 6-year-old children, or were developed specifically for this study. Six ability domains were hypothesized:

---

[1] Supported in part by the National Institute of Mental Health Grant No. MH 08666-01: Differential Abilities of Six-Year-Old Children. Appreciation is expressed to the Los Alamitos Elementary School District Board of Trustees, administrators, kindergarten teachers, children, and parents who cooperated with the study; to Paul Sundstrom, for invaluable psychometric assistance; and to Michi Yomogida for painstaking editorial asistance. Data were processed by the Western Data Processing Center, University of California at Los Angeles.

1. *Visual memory for figural units* (MFU). Test 1, Cube Test (Hildreth & Pintner, 1937). *S* taps cubes in same sequence as examiner. Test 2, Picture Memory (*S* names objects previously seen, now screened). Test 3, Memory for Patterns (*S* points to previously seen pattern when mixed with others).

2. *Auditory memory for symbolic units* (MSU). Test 4, WISC Digit Span, forward (Wechsler, 1949). Test 5, Auditory Letter Span. Test 6, Memory for Sentences, taken from Binet (Terman & Merrill, 1937) with added items. Test 7, WISC Digit Span, backward (Wechsler, 1949).

3. *Convergent semantic production* (NM). Test 8, WISC Comprehension (Wechsler, 1949). Test 9, WISC Vocabulary (Wechsler, 1949). Test 10, Action-Agent, Convergent (Stutsman, 1931) with more difficult items added by McCartin and Meyers (1966).

4. *Divergent production of semantic units—ideational fluency* (DMU). *S* must produce a variety of acceptable responses in each task. Test 11, Action-Agent, Divergent, modified from McMartin and Meyers (1966). Test 12, Monroe Language Classification (Monroe, 1935). Test 13, Utility Test, adapted from Guilford and Christensen (1956). Test 14, Thing Categories, adapted from French, Ecstrom, and Leighton (1963).

5. *Evaluation of figural units* (EFU). Speedy accuracy is required in identifying or matching figural materials. Test 15, SRA-PMA, Identical Pictures (Thurstone & Thurstone, 1953). Test 16, SRA-PMA, Identical Figures (Thurstone & Thurstone, 1953). Test 17, ETS Identical Pictures (French et al, 1963).

6. *Evaluation of symbolic units* (ESU). Like 5 but with symbolic materials. Test 18, Word Comparison (Harrison & Stroud, 1956). Test 19, Number Comparison. Test 20, Finding 0's.

SUBJECTS AND PROCEDURES

The *S*'s were 100 anglo-white children, 50 boys and 50 girls, in five public elementary schools from a middle-class residential suburb of Long Beach, California. CA's ranged from 71 to 75 months with a mean of 72.7 months. The mean IQ based on three WISC subtests used was 115.5. The battery required over an hour of individual testing, given in two, sometimes three or more sessions.

RESULTS

Raw score data, none requiring data transformation, were employed. Intercorrelations yielded 12 principal component factors by computer program. These were given several rotations. Information presented here is based on a varimax orthogonal rotation (Kaiser, 1959) of seven factors (the number of hypotheses plus one). Other rotational options, each based on minor logical differences, were employed to determine whether the resulting structure would differ in psychologically meaningful ways. With one exception to be discussed the options did not change the interpretation. The analysis is presented as a whole in Table 1, with the first six factors (A–F) rearranged in an order most likely to represent the hypotheses.

1-MFU. Factor A is tentatively identified as support for the hypothesized factor, as it saturates two of the variables intended, together with digits forward, which was intended for MSU.

Test 1	Knox cube tapping	.63

Test 2	Picture memory	.09 (.32 on F)
Test 3	Memory for patterns	.35 (.39 on F)
Test 4	Digits forward, designed for MFU	.43

2-MSU. Tests 4, 5, and 6 load well on Factor B which may be identified as the sought-after MSU. Digits forward divided its loading.

Test 4	Digits forward	.47 (.43 on A, MFU)
Test 5	Auditory letters	.66
Test 6	Memory for sentences	.64
Test 7	Digits backward	.17 (.33 on E)

3-NM. Factor C appears to support this hypothesis. One variable had a second loading.

TABLE 1

VARIMAX ROTATED FACTOR LOADINGS

	A	B	C	D	E	F	G	h^2
1. MFU, Visual memory for figural units								
1. Cube test	63^a	10	08	06	12	06	−01	44
2. Picture memory	09	03	17	12	08	32	01	26
3. Memory for patterns	35^a	14	11	07	25	39^a	09	42
2. MSU, Auditory memory for symbolic units								
4. WISC Digits Forward	43^a	47^a	03	05	05	18	01	47
5. Auditory Letter Span	00	66^a	−08	−03	03	−04	00	47
6. Memory for Sentences	22	64^a	27	12	16	08	−03	59
7. WISC Digits Backward	29	17	20	30	33	04	18	47
3. NM, Convergent semantic production								
8. WISC Comprehension	12	04	69^a	19	01	02	05	55
9. WISC Vocabulary	−01	07	71^a	27	19	24	−14	71
10. Action-Agent Convergent	08	−04	48^a	16	06	11	39^a	46
4. DMU, Divergent production of semantic units								
11. Action-Agent Divergent	01	−08	02	59^a	00	25	−17	46
12. Monroe Language Classification	00	25	22	59^a	30	14	−03	60
13. Uses	07	−01	22	66^a	09	06	06	54
14. Thing Categories	09	04	18	57^a	08	−13	16	43
5. EFU, Evaluation of figural units								
15. PMA Identical Pictures	19	09	14	18	70^a	−25	−10	68
16. PMA Identical Figures	14	09	29	20	64^a	−04	−32	70
17. ETS Identical Pictures	19	11	02	19	81^a	−06	−03	78
6. ESU, Evaluation of symbolic units								
18. Word comparison	22	01	05	−05	73^a	20	03	68
19. Number comparison	−01	−03	−04	08	52^a	23	15	38
20. Findings 0's	−16	−08	05	00	58^a	22	09	47

Note.—Decimals have been omitted. h^2 (communality) Based on the 12 extracted factors; however, only 7 factors were rotated.

[a] Indicates loaded factors.

Test 8	WISC comprehension	.69
Test 9	WISC vocabulary	.71
Test 10	Action agent, convergent	.48 (.39 on G, a singlet)

4-DMU. This appears to be represented in Factor D. All four tests loaded as predicted to support the hypothesis.

Test 11	Action agent, divergent	.59
Test 12	Monroe language classification	.59
Test 13	Uses	.66
Test 14	Thing Categories	.57

5-EFU. Easily identified as evidence for EFU, Factor E shows strong loadings on all hypothesized variables, but also on the ESU tests.

Test 15	SRA PMA identical pictures	.70
Test 16	SRA PMA identical figures	.64
Test 17	ETS identical pictures	.81
Test 18	Word comparison, designed for ESU	.73
Test 19	Number comparison, designed for ESU	.52
Test 20	Finding 0's, designed for ESU	.58

6-ESU. Table 1 shows that no strong factor loads on Tests 18, 19, and 20, designed for ESU, except of course for the EFU factor. When rotation of all 12 factors was performed, the picture changed somewhat. While these three tests continued to load on EFU, a good Factor "F" appeared, with substantial saturations of these variables, sufficient to suggest the ESU factor might be represented.

DISCUSSION

Well-defined factors appear to have emerged and to represent evidence for four of the six hypothesized abilities: MSU, auditory memory for symbolic units; NM, convergent semantic production; DMU, divergent production of semantic units; and EFU, evaluation of figural units (perceptual speed). MFU, visual memory for figural units, was uncertain and tentative. ESU, evaluation of symbolic units, was hypothesized with perceptual tasks of letters, numbers, and words, but could not be separated from EFU except at the caprice of a rotational option.

The demonstration of a short-term memory factor orthogonal to other abilities replicates similar work at the developmental level in question (Kebbon, 1964; Kelley, 1928; Loeffler, 1963; and McCarthy & Kirk, 1964 in some but not all of their analyses). That two or more short-term memory factors may exist has some support, such being found in retarded of MA 4 by Carlson, Meyers, and Dingman (1965) in which the visual factor was unclear, just as it was in the present study, though the separation from the auditory was certain enough. Digits backward did not join the MSU factor, but its best saturation, with DMU, is no surprise considering the nature of the backwards task.

The NM or convergent semantic production factor appears to be a firm member of the repertory of 6-year-old abilities. This is an aspect of the traditional general-verbal ability celebrated in age scales, and replicates the factorial work by Loeffler (1963) and McCartin and Meyers (1966). An issue of factor name might arise. McCartin and Meyers had not been able to make a very sharp factorial distinction between cognition and convergent operations (discussed above). It is possible that at this age the two cannot be separated in practical testing. The present tasks were judged to require more of the child's formulation of explanations than of comprehending verbal input, so the "N" or convergent label was preferred. The question of how much difficulty of "C" or cognition entered in the tasks is not presently answerable. As to the absence of a third, or products, term in the lexicon of the structure of intellect, (as U for units, C for classes, etc.), this was intentionally avoided since the children's responses to Tests 8, 9, and 10 had to range from single words to sentences.

DMU or divergent semantic units also appears to hold a secure position in the adaptation repertory of this age. This "ideational fluency" factor was obtained with some of the tests used by McCartin and Meyers (1966) and some new ones, and is apparently the same factor as that given different names by others (Bereiter, 1961; Loeffler, 1963; and McCarthy & Kirk, 1964).

Evaluation of figural units or EFU is a good factor, the results here supporting the investigations of Kebbon (1963), Loeffler (1963), Meyers et al (1964), Meyers et al (1962), Stott and Ball (1963), and Thurstone and Thurstone (1953). This is the "perceptual speed" factor, under this and other names. The factor also captured the variance of the ESU tests, with the implication that children this age still must react to visual symbols of language as perceptual tasks. While this study produced a separation of MSU from MFU in the memory domain, that difference was due in part to the difference in sensory input channel, a difference not present in the ESU-EFU contrast.

In summary, the present study has confirmed and extended the knowledge of primary ability factors in children of 6. Considering the present evidence in the light of other reports, the hypotheses of MSU, NM, DMU, and EFU appear to be confirmed now beyond reasonable question, with the possibility that MFU may be yet developed.

REFERENCES

BEREITER, C. Fluency abilities in pre-school children. *Journal of Genetic Psychology*, 1961, **98**, 47–48.

CARLSON, D. C., MEYERS, C. E., & DINGMAN, H. F. An investigation of language, memory, and perceptual abilities at mental age four. Unpublished manuscript, University of Southern California, 1965.

FRENCH J. W., ECSTROM, R. B., & LEIGHTON, A. B. *Manual for Kit of Reference Test for Cognitive Factors*. Educational Testing Service, 1963.

GUILFORD, J. P. Three faces of intellect. *American Psychologist*, 1959, **14**, 46–79.

GUILFORD, J. P., & CHRISTENSEN, P. R. A factor analytic study of verbal fluency. *Reports from the Psychological Laboratory*, No. 17, Los Angeles: University of Southern California, 1956.

GUILFORD, J. P., & HOEPFNER, R. Current summary of structure-of-intellect factors and suggested tests. *Reports from the Psychological Laboratory*, No. 30, Los Angeles: University of Southern California, 1963.

HARRISON, M. L., & STROUD, J. B. *Reading readiness profiles*. Boston: Houghton Mifflin, 1956.

HILDRETH, G. H., & PINTNER, R. *Manual of Directions for Pintner-Patterson Performance Tests*, Short Scale. Bureau of Publications, Teachers College, Columbia University, New York, 1937.

KAISER, H. F. Computer program for varimax rotation in factor analysis, *Educational and Psychological Measurement*, 1959, **19**, 413–420.

KEBBON, L. Ability structure and special defects in the mentally retarded. International Copenhagen Congress on the Scientific Study of Mental Retardation, August 7–14, 1964. (Abstract)

KELLEY, T. L. *Crossroads in the mind of man*. Stanford: Stanford University Press, 1928.

LOEFFLER, F. J. An extension and partial replication of Meyers' primary mental abilities of MA6. Paper delivered at Society for Research in Child Development Meeting, Berkeley, 1963.

MCCARTHY, W., & KIRK, S. A. *Illinois Test of Psycholinguistic Ability, Manual*. Madison, Wisconsin: University of Illinois, 1961.

McCARTHY, J. J., & KIRK, S. A. *The construction, standardization, and statistical characteristics of the Illinois Test of Psycholinguistic Abilities.* Urbana: Institute for Exceptional Children, 1964.

McCARTIN, SR. R. A., & MEYERS, C. E. An exploration of six semantic factors at first grade. *Multivariate Behavioral Research,* 1966, in press.

MEYERS, C. E., & DINGMAN, H. F. Factor analysis and structure of intellect applied to mental retardation. *Mental Retardation Abstracts,* American Association of Mental Deficiency Documentation Service, Columbus, Ohio, 1965, **2,** 116–127.

MEYERS, C. E., DINGMAN, H. F., ORPET, R. E., SITKEI, E. G., & WATTS, C. A. Four ability-factor hypotheses at three preliterate levels in normal and retarded children. *Monographs of the Society for Research in Child Development,* 1964, **29** (5, Whole No. 96).

MEYERS, C. E., ORPET, R. E., ATTWELL, A. A., & DINGMAN, H. F. Primary mental abilities at mental age six. *Monographs of the Society for Research in Child Development,* 1962, **27**(1, Whole No. 82).

MONROE, M. *Reading aptitude tests.* Boston: Houghton Mifflin, 1935.

STOTT, L., & BALL, R. S. *Evaluation of infant and pre-school mental tests.* Detroit: The Merrill Palmer Institute, 1963.

STUTSMAN, R. *Mental measurement of preschool children.* New York: World Book Company, 1931.

TERMAN, L. M., & MERRILL, M. E. *Measuring intelligence: A guide to the administration of the New Revised Stanford-Binet Tests of Intelligence.* Boston: Houghton Mifflin, 1937.

THURSTONE, L. L., & THURSTONE, T. G. *SRA Primary Abilities for Ages 5–7.* Chicago: Science Research Associates, 1953.

WECHSLER, D. *Wechsler Intelligence Scale for Children,* New York: Psychological Corporation, 1949.

(Received October 29, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 6, 347–349

METHODOLOGICAL NOTE ON FEELING-OF-KNOWING EXPERIMENTS[1]

J. T. HART
University of California, Irvine

Further investigations of the ability of Ss to accurately predict memory contents when they are unable to retrieve those contents, a process termed the memory-monitoring or MEMO process, are reported. 2 possible sources of error in the earlier experiments were examined—either source could have falsely inflated the MEMO accuracy scores. The replication findings show that it is important to instruct Ss to guess even when they believe that they do not know the correct answer. The instruction to guess forces overcautious Ss to provide answers that they would otherwise withhold. Answers guessed correctly can be eliminated before MEMO accuracy is evaluated, thus overestimates of MEMO accuracy are avoided.

Is it possible for people to know what is in memory even when they are unable to retrieve the memories? Evidence provided in two previous experiments (Hart, 1965) suggests that they can. The accuracy of feeling-of-knowing experiences was assessed by applying a recall-judgment-recognition (RJR) paradigm, in which subjects (*S*s) first took a general information test of recall and were instructed to try to predict recognition failures and successes for items they were unable to answer by referring to their feelings of knowing about the sought-after items. After taking the recall test and making feeling-of-knowing judgments, (FK) (yes) or $\overline{\text{FK}}$ (no),[2] *S*s were given a recognition test. Usually *S*s were able to predict accurately which unrecalled items they would recognize and which they would not recognize.

The intervening process yielding FK and $\overline{\text{FK}}$ judgments is termed the memory-monitoring (MEMO) process. An accurate MEMO process could be useful to a fallible storage and retrieval system as an indicator of when to continue and when to stop retrieval efforts.

There is some possibility that the experimental procedure used in the previous experiments overestimated MEMO accuracy. The follow-up experiments reported in this article attempt to eliminate sources of accuracy inflation.

METHOD

Experiment I

In the earlier experiments *S*s made feeling-of-knowing (FOK) judgments for those items they could not answer. The instructions specifically told them to make FOK judgments "for all those items you must leave blank." This instruction assumes that items will be left blank only when *S*s believe they have not remembered the correct answers. Otherwise it does not make sense to talk about the MEMO process as an indicator of what is in storage at times when retrieval is not successful, and to use the prediction of recognition successes and failures to assess MEMO accuracy. Interviews with some of the *S*s after the completion of the experiments revealed that this key assumption about the meaning of blanks in the test of recall was not entirely warranted.

Some *S*s reported leaving the answer spaces blank, and making FK judgments, at times when they strongly believed they had already retrieved the correct answer. This

---

[1] The author gratefully acknowledges the assistance received from Albert Hastorf, Leonard Horowitz and Karl Pribram.

[2] See Hart (1965) for a complete explanation of the terms FK Hit and Miss and $\overline{\text{FK}}$ Hit and Miss; they correspond to the psychophysical terms True Positive and False Positive, and True Negative and False Negative.

practice arose because, occasionally, *S* retrieved more than one answer to a question (one of which he felt would be correct) but was unable to decide among the several alternatives. This could have produced a false inflation of the FK hits and the comparison of inflated FK hits to the $\overline{\text{FK}}$ misses would lead to a false conclusion about MEMO accuracy. Inflation of an FK hit score would occur any time the alternatives on the recognition test did not exactly overlap the alternatives *S* had provided for himself during the test of recall. In effect, the experimenter would help *S* to choose between the alternatives.

To evaluate this possible source of error Experiment I was conducted using a modified set of instructions. The modified instructions told *S*s to write several answers if they were unable to decide upon one. By using this instruction, all questions receiving more than one answer could be eliminated prior to the MEMO accuracy analyses. Aside from these slightly modified instructions, the procedure and materials used were the same as those of the previous Experiment II (Hart, 1965). The 18 *S*s were drawn from the same source, a general psychology course at Stanford University.

RESULTS

The results show that the instruction about multiple answers was effective. Twice as many errors were recorded in this experiment as in the earlier study, and there were fewer FK items and fewer blanks. This is the pattern of results to be expected if *S*s followed instructions, since those multiple-answer items formerly left blank and marked as FK's would be differently categorized.

But the altered instructions did not influence MEMO accuracy scores. The mean of the FK hits is .70, $\overline{\text{FK}}$ misses .39. A test of the FK and $\overline{\text{FK}}$ proportions of correct responses is highly significant ($t = 5.25$, $df = 16$, $p < .01$, one-tailed).

For a complete presentation of the rationale behind this comparison, see Hart (1965). The basic point of the FK–$\overline{\text{FK}}$ test is that, if FOK experiences are accurate indicators of memory storage, *S*s should recognize more items they feel they know (FK items) than items they feel they don't know ($\overline{\text{FK}}$ items). The comparison of proportions correct for FK and $\overline{\text{FK}}$ items is a *relative* accuracy score and, as such, is independent of the ease or difficulty of the recognition test (unlike the absolute number of FK and $\overline{\text{FK}}$ prediction hits).

When multiple answers are given to questions, the answers are as often wrong as they are right. Consequently, withheld multiple answers do not provide a significant source of error in FK scores.

METHOD

Experiment II

Not until Experiment I had been completed was yet another possible source of error recognized. Postexperiment interviews with selected *S*s from Experiment I revealed that the *S*s appeared to differ significantly in their willingness to write down a guess on the answer sheets during the test of recall. Some *S*s were extremely cautious and would enter an answer only when they felt completely sure of its correctness. Others were willing to write anything that seemed better than a sheer guess.

Overcautious answering by many *S*s could lead to a false increase in the number of FK hits, since *S*s would leave some questions unanswered and give high FOK ratings at times when they were nearly, but not entirely, certain of the correct answer.

The simple way to eliminate this source of error is to require *S*s to construct an answer, even if they must guess, for *every* question. A FOK rating is made only on those questions that *S*s believe they have not answered correctly. Since none of the items answered correctly on the test of recall are included in the evaluation of MEMO accuracy, the instruction to guess should prevent overcautious withholding of correct answers and inflation of FK hit scores.

Twenty-five questions were added to the recall and recognition tests.[3] These questions were added in case the instruction to guess resulted in a sizeable decrease in the absolute number of FK judgments made by the *S*s. With 100 questions it was hoped that enough FK items would remain to yield some evaluation of MEMO accuracy. Fourteen students from an introductory general psychology course served as *S*s.

---

[3] Some of these additional questions were adapted from a 200-item questionnaire developed by Murdock (1963). The author is indebted to B. B. Murdock for kindly sup-

RESULTS

The FK scores obtained in Experiment II were the same as those in Experiment I. With the addition of 25 questions the number of FK responses should have increased but didn't. Apparently, instructions to guess eliminated some of the items that had formerly received high FK ratings because the *S*s were nearly sure, but not certain, of the correct answers.

Further evidence that the FK scores have been somewhat reduced is that the mean of the FK hits dropped from .70 in Experiment 1 to .52 in this Experiment, ($t = 3.6$, $p < .01$); $\overline{\text{FK}}$ means were not changed.

However, even with the reduction in the inflated FK hit score, the comparison of the proportion of responses correct for FK and $\overline{\text{FK}}$ items still yields a significant result in the predicted direction ($t = 2.46$, $df = 13$, $p < .025$, one-tailed). The MEMO process is accurate, but less accurate than had appeared in the earlier experiments.

The results of the two experiments may be succinctly summarized. The presence of multiple alternatives to a few questions does not produce an inflated estimate of the number of true FK positives. No special provision need be made for handling multiple alternatives. Overcautious withholding of answers, however, does constitute a significant source of FK inflation. Consequently, all experiments employing the RJR paradigm should instruct *S*s to provide answers to every recall question, even if they must guess.

REFERENCES

HART, J. T. Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 1965, 56, 208–216.

MURDOCK, B. B. An analysis of the recognition process. In C. N. Cofer & B. S. Musgrave (Eds.), *Verbal behavior and learning*. New York: McGraw-Hill, 1963. Pp. 10–22.

(Received November 5, 1965)

---

plying copies of his recall and recognition tests.

Journal of Educational Psychology
1966, Vol. 57, No. 6, 350–358

EFFECT OF SELECTED METHODOLOGICAL CONDITIONS ON DEPENDENT MEASURES TAKEN AFTER CLASSROOM EXPERIMENTATION[1]

WILLIAM L. GOODWIN[2]
University of Wisconsin

The necessity of controlling the influence of selected extraneous variables upon data collected after classroom experimentation was investigated. The experimental units were 64 6th-grade classes, grouped into 4 strata according to previous arithmetic achievement. Within each stratum, classes were randomly assigned to 1 of the 16 experimental treatments generated by a 2^4 factorial design using 4 independent variables: experimental atmosphere and absence of same; notice of test and no notice; teacher administration and outside administration; and teacher scoring and outside scoring. Resulting class means on subtests of a standardized arithmetic achievement test were subjected to analyses of variance. Significantly higher ($p < .10$) class means were found for those classes whose regular teacher administered the test. Several interactions were also significant.

Controlled experimentation and other research activities are becoming increasingly prevalent in school settings. There exist a number of extraneous variables that may influence the precision of the data collected after classroom experimentation. Many researchers use the teacher to administer treatments and tests or to assist in varying ways; in effect, the teacher is cast in the role of a subexperimenter. Only limited systematic observation and measurement of the effect of subexperimenters on experimental results has occurred.

The researcher must decide whether or not to inform teachers that an experiment is in progress. It is implied by Orne (1962) that some teachers, even when uninformed, will perceive that an experiment is underway and will speculate about the hypotheses under investigation. If, on the other hand, the experimenter informs teachers of the experiment, he may subtly bias them in directions favorable to his opinions, as Rosenthal suggests (1963, 1965). In addition, teachers and pupils, being informed of the experiment, may perform unnaturally well because of the Hawthorne Effect, thereby jeopardizing the external validity of the results (Cook, 1962; Mayo, 1945a, 1945b; Roethlisberger, 1941; Roethlisberger & Dickson, 1941).

Notice of test date is potentially more reactive where pupil progress on standardized achievement tests is used as an indication of teacher merit (Douglass, 1935; Simpson, 1947; Thorndike & Hagen, 1961). The researcher, however, working with the school administrators, can schedule the testing so that any desired period of notice can be achieved.

Utilizing teachers to give and score tests at the end of an experiment is an alluring course of action; it is convenient and inexpensive, and allows pupils to react naturally and optimally, as their regular teacher is the test administrator. However, teachers may be biased in favor of, and have rapport established with, their pupils to varying degrees. In some cases, bias is a natural and desirable phenomenon, but an experiment is not one of these cases, as such bias would result in increased uncontrolled variance. (No less important, although not investigated in this study, is the potential bias of outside test administrators.) From the Hopkins and Lefever study (1965), one would expect no differences between pupil performance on teacher-administered and outside-administered tests. Phillips and Weathers (1958) suggest that errors made by teachers in test scoring would raise grades as often as lowering them. No literature is directly pertinent to the question of the differential effects of scoring by teachers and outsiders in the case at hand, namely because the scoring is a routine procedure involving multiple-choice responses and a scoring key with few judgments by the scorer.

---

[1] The research reported herein was performed pursuant to a contract (OE 5-10-154, Center No. C-03) with the United States Office of Education, Department of Health, Education, and Welfare (Goodwin, 1965). The author also wishes to acknowledge the suggestions and assistance of Herbert J. Klausmeier and Julian C. Stanley.

[2] Now at Bucknell University.

Specifically, the problem to be researched has four facets: What is the effect of varying conditions of experimental atmosphere, notice of testing, test administrator (teacher or outsider), and test scorer (teacher or outsider) upon student performance as measured by test results? The importance of these methodological considerations for research in education and educational psychology is readily apparent.

METHOD

Subjects

The experimental units were 64 sixth-grade classrooms in a large Midwestern city, selected from among 87 schools which had one or more classes containing at least 15 second semester sixth-graders. One class was selected from each school having more than one such class by using a table of random numbers. It was necessary to include only one class per school because of the possible reactive or contaminating nature of the experimental treatments.

To reduce random variability in the design, a stratified random sampling procedure was used. The 87 classes were ranked on the basis of previous arithmetic achievement (class means on the *Iowa Test of Basic Skills* given earlier in the school year), and subsequently grouped by fourths, from highest to lowest. Within each of the four resulting strata, classes were randomly assigned to one of 16 experimental treatments. The resulting 64 classes were taught by 38 male and 26 female teachers; in all, 1657 students' performances were analyzed. (Only arithmetic scores of pupils enrolled in the second semester of sixth grade were used in determining strata and in reaching conclusions, although all the pupils in each class were tested in order to achieve realism and credibility in all treatments.)

Procedure

The 16 experimental treatments were the combinations generated by a 2^4 factorial design using the following independent variables in connection with a recent arithmetic achievement test as a response measure:

1. Experimental atmosphere (+) and absence of the same (−);
2. Notice of test date (+) and no notice (−);
3. Testing by regular teacher (+) and testing by outsider (−); and
4. Scoring by regular teacher (+) and scoring by outsider (−).

Treatment under Variable 1 was effected by letter from the office of the school system's director of research. Experimental units under the experimental-atmosphere condition were informed by mail 14 days before the testing date that they were in an experiment. Units not under the experimental-atmosphere condition were told when notified of the test date that they had been randomly selected to collect normative data for a new standardized test.

Notice of test date was effected by mail 14 days prior to test date. Under the no notice condition, teachers were not informed of the testing until the Friday preceding the Monday test date.

The test-administrator variable was introduced by sending copies of the test and manual to the teacher administrators whenever their class received notice of the testing. Outside administrators (graduate and advanced undergraduate students in measurement courses) were given the test packets by their college instructors after being randomly assigned to classes. All administrators (teacher and outside) also received detailed written instructions on how to prepare for administering the test; neither group was contacted personally by the experimenter or his assistants to discuss proper test procedures. The instructions received caused the outside administrators to believe either that an experiment was or was not in progress, according to the experimental treatment the particular class was receiving.

The test-scoring variable was implemented by leaving the exams with half the teachers for scoring within four days. The other tests were collected and scored by four outsiders whose orientation in regards to experimental atmosphere was identical to that of the teachers whose exams were scored (two of the scorers believed that the tests resulted from an experiment while the other two believed that normative data were being collected). All tests were subsequently rescored to determine the accuracy of the initial scoring.

Four related dependent variables were investigated: raw scores in arithmetic computations, concepts, applications, and a total or average score. These four quantities are yielded by the *Stanford Arithmetic Achievement Test*, Intermediate II. The Computations subtest uses 39 standard or drill-type items primarily concerned with fundamental arithmetic processes. The Concepts subtest concerns itself with the ideas or concepts behind the fundamental arithmetic processes, consists of 32 items, and is more verbal than the Computations subtest. The third subtest, Applications, contains 39 word problems in which the pupil must interpret graphs, compute areas, figure sales tax, etc. This test was copyrighted in 1964 and was unfamiliar to the teachers in the selected school system.

RESULTS AND DISCUSSION

Resulting class means on each of the four dependent variables were subjected to a 4×2^4 analysis of variance. The error term initially was composed of all four-factor interactions and the single five-factor interaction ($df = 16$); these higher order interactions were assumed to be estimates of σ^2. It was decided a priori to use this error term to test the remaining three-factor interactions that involved the leveling variable before pooling further. The sums of squares and degrees of freedom of statistically nonsignificant interactions were to be included in the final error term; on the other hand, a significant interaction, or even one approaching significance, certainly could not be assumed to be zero or small, could not be considered an estimate of σ^2, and was not to be pooled (Green & Tukey, 1960). Three-factor interactions that included the leveling variable were tested using the initial error term, and any interaction whose F ratio exceeded twice the 50% point of the F distribution with the corresponding degrees of freedom was not pooled. The nonsignificant interactions were assumed to be estimates of σ^2 and were pooled to form the final error term. Only one three-factor interaction, Test Administrator × Scorer × Previous Achievement, fell into the no-pool category on three of the four dependent variables; on the fourth subtest, Computations, no F ratio reached the critical value of 1.65.

Of the main effects, only previous arithmetic achievement tested highly significant for each of the dependent variables ($p < .001$). Notice of test reached significance ($p < .05$) in the case of the Concepts subtest; classes receiving notice of the test (+) averaged 15.51 while those classes receiving no notice (−) averaged 14.64 in raw score units (a difference of approximately 2 months in grade placement units). None of the other main effects were significant; the relevant means are presented in Table 1.

Interestingly, test notice, although significant for the Concepts subtest, apparently had a complete lack of effect in the case of the Computations subtest. Examination of the test items involved suggests a plausible explanation. The computation items are rou-

TABLE 1

RAW SCORES ON SUBTESTS OF STANFORD ARITHMETIC ACHIEVEMENT TEST BY INDEPENDENT VARIABLE

Independent variable		Subtest						Average	
		Computation		Concepts		Applications			
		X̄	s	X̄	s	X̄	s	X̄	s
Experimental atmosphere	+	18.12	3.95	15.30	3.59	20.58	4.46	18.00	3.85
	−	17.88	4.21	14.85	3.42	20.15	4.44	17.63	3.84
Notice of testing	+	17.95	4.45	15.51	3.72	20.61	4.73	18.02	4.17
	−	18.05	3.67	14.64	3.24	20.12	4.15	17.60	3.48
Test administrator	+	18.54	4.08	15.44	3.43	20.70	4.45	18.23	3.84
	−	17.46	3.90	14.72	3.56	20.03	4.43	17.40	3.81
Test scorer	+	18.23	3.79	15.11	2.97	20.58	3.42	17.97	3.28
	−	17.77	4.24	15.04	3.99	20.15	5.27	17.65	4.34

tine, drill-type problems. Normally, students have attained their ability to do this type of problem over several years of frequent practice. An inordinate emphasis practicing similar items would be necessary to bring about any appreciable gain in the students' performance on the task. However, the problems in the Concepts subtest are more of the "aha" variety, that is, extremely puzzling when first encountered but relatively routine after a short discussion of the concepts involved. It is possible that teachers who received notice of the testing read over the test items. With no conscious motivation to aid their pupils on the tests, they may have been attracted by some of the concept (and application) problems and subsequently may have discussed that type of problem with their classes. Pupils in no-notice classes would not have had a similar opportunity to learn of the concepts involved in the test items.

The entries in Table 1 consistently favor the + experimental atmosphere treatment, but the associated F ratios take values far from significance. A critical difference between this study and those mentioned above which contend that experimental atmosphere is a potent variable is that in the current investigation the full burden of conveying or administering the experimental atmosphere condition fell upon a short, and in some respects innocuous, paragraph of instructions. This simulated experimental atmosphere is noticeably different from the daily interaction between experimenter and subject that took place in the Hawthorne Western Electric plant. The obvious possibility should not be overlooked that experimental atmosphere may, in actuality, have little effect on teachers in many situations.

When teachers administered the test, resultant class means were higher by nearly a raw score point, representing 2 months' achievement, than were means of classes in which outsiders gave the test (Table 1). The average F ratio for this main effect is significant at approximately .10, considerably different from the .05 level used heretofore. The consistency of the effect across all subtests, however, lends support to its claim as a true, rather than a chance difference. If this is a true difference, its cause might be due to several factors or some combination of them. Rice (1897), Lowell (1919), and Traxler (1951) suggested that indirect hints given by the teacher during the test might aid pupils to obtain higher scores. Tacit understandings and rapport between teacher and pupil might be equally important considerations. Anxiety may reach a higher level when outsiders administer the test, especially for younger pupils; in some cases, this anxiety may impair the pupil's performance. In any event, the investigator should consider alternatives that permit testing classes under nearly identical conditions; the use of television (Hopkins & Lefever, 1965) or prerecorded instructions offer promise of reducing the uncontrolled variance inherent in using several test administrators.

From Table 1, it would appear that the question of possible differential effects when tests are scored by teachers and outsiders (and when the scoring is a routine procedure involving few judgments by the scorer) is not a critical one; however, discussions below of the significant interactions suggest otherwise. An analysis of the percent-error rates made by the teacher-scorers demonstrated that their average error rate was comparable to that of the outside scorers. The errors made by the teachers were as likely to raise as lower grade placements, supporting an earlier finding of Phillips and Weathers (1958). The wide variation in error rates of the scorers, both teachers and outsiders, suggests that it would be desirable to use a few highly competent scorers, thereby reducing error variance resulting from inaccurate scoring.

Two first-order interactions reached significance, each on three of the four dependent variables. The means generated by the interaction of Experimental Atmosphere × Test Notice are presented in Table 2.[3] The source of this significance was a crossover resulting from a high mean raw score for the ++ treatment combination, low means for the +− and −+ combinations, and an intermediate mean raw score for the −− combination. This interaction not only highlights the differential effectiveness of an aura or atmosphere of experimentation when combined with notice of an approaching test, but also raises the question of why the −− cell classes performed so well.

The other significant two-factor interaction was Test Scorer × Previous Arithmetic Achievement, the stratifying variable; relevant means are presented in Table 3. As can be noted in the table and Figure 1, a marked crossover occurs. The source of this significance is the higher means for outside scorers on Stratum 1 classes' tests (high previous achievers), and the reversal of this situation for Strata 3 and 4 (i.e., higher raw-score means for teacher-scorers). It suggests that the teachers in the upper stratum classes put more emphasis on motivating and/or preparing their pupils for the test when they (the teachers) knew that it would be scored by outsiders than when they were to score it themselves. On the other hand, teachers in the lower strata did not increase their preparation and/or motivation efforts

---

[3] In this table and all subsequent tables, algebraic signs will be used to indicate treatment interactions in accordance with the treatment definitions given in the procedures section. The first sign will refer to the first term of the interaction, the second sign to the second term, etc. For example, in Table 2 the significant interaction is listed as Experimental Atmosphere × Test Notice; thus, + + would refer to + experimental atmosphere and + notice of testing, − + would refer to − experimental atmosphere and + notice of testing, etc.

TABLE 2

RAW SCORES ON SUBTESTS OF STANFORD ARITHMETIC ACHIEVEMENT TEST: EXPERIMENTAL ATMOSPHERE × TEST NOTICE TREATMENT COMBINATION

Experimental Atmosphere × Test Notice treatment combination	Subtest						Average	
	Computation		Concepts		Applications			
	X̄	s	X̄	s	X̄	s	X̄	s
+ +	18.85	4.22	16.10	3.56	21.41	4.53	18.79	3.91
+ −	17.38	3.64	14.49	3.55	19.75	4.37	17.21	3.73
− +	17.05	4.63	14.92	3.89	19.82	4.92	17.26	4.42
− −	18.71	3.70	14.79	3.01	20.49	4.02	17.99	3.28
F ratio	5.55*		3.39		6.97*		7.17*	

* $p < .05$.

when they knew the test would be scored by outsiders. The average differences on the three subtests favored the teacher-scorers by .69, 1.14, and 1.36 raw-score points (or approximately 1, 2, and 3 months' school achievement) for Strata 2, 3, and 4, respectively. Only the fact that the opposite situation was true in Stratum 1 (where the mean of the tests scored by outsiders exceeded that of the tests scored by teachers by 1.92 raw-score points, over 4 months' achievement) kept the main effect of test scorer from reaching significance.

Statistical significance was again reached when this two-factor inter-

TABLE 3

RAW SCORES ON SUBTESTS OF STANFORD ARITHMETIC ACHIEVEMENT TEST: TEST SCORER × PREVIOUS ARITHMETIC ACHIEVEMENT (STRATUM) COMBINATION

Previous arithmetic achievement (stratum)		Subtest						Average	
		Computation		Concepts		Applications			
		TS	OS	TS	OS	TS	OS	TS	OS
1 (High)	X̄	20.79	22.19	17.66	19.74	23.76	26.05	20.74	22.66
	s	2.69	2.69	1.64	2.26	1.62	2.13	1.89	2.04
2	X̄	20.40	18.69	17.15	16.53	22.44	22.69	19.99	19.30
	s	2.13	2.75	1.41	1.57	1.42	1.69	1.41	1.66
3	X̄	18.01	16.97	14.42	13.59	20.10	18.53	17.51	16.37
	s	3.61	2.91	1.43	2.00	2.18	2.66	2.27	2.08
4 (Low)	X̄	13.73	13.22	11.20	10.31	16.02	13.33	13.65	12.29
	s	1.89	2.96	1.44	1.79	1.67	2.09	1.45	2.32
F ratio		1.01		3.26*		6.09**		3.11*	

Note.—Abbreviations are: TS, Teacher Scorer; OS, Outside Scorer.

* $p < .05$.

** $p < .01$.

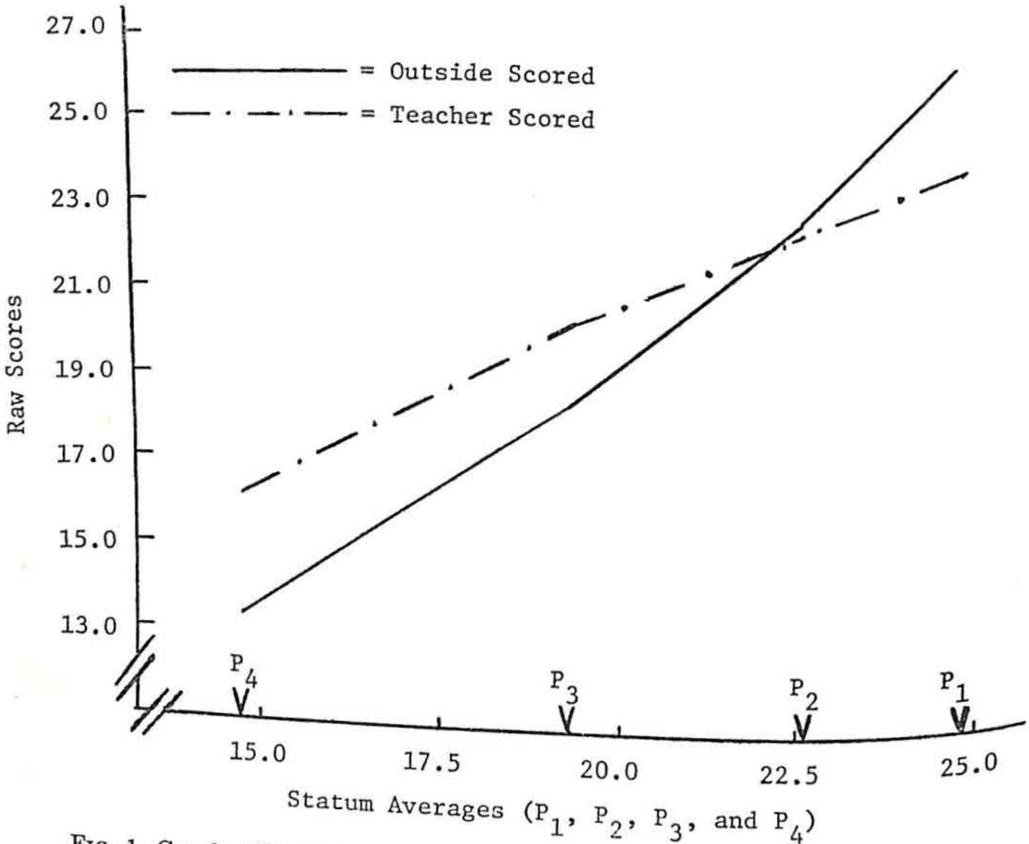


FIG. 1. Graph of Test Scorer × Previous Arithmetic Achievement interaction on applications subtest.

action was paired with the effect due to test administrator, that is, the Test Administrator × Test Scorer × Previous Achievement Level interaction. As shown in Table 4, some highly disparate means contributed to the significance found. Inspecting means for the third and fourth strata for the four possible combinations of the administrator and scorer variables, differences of 1.03 to 4.02 are found (which, converted, range from 2 months to over 7 months' achievement in grade-equivalent units) favoring the + + treatment combination over the + − cell. The reverse situation is true for Stratum 1, with the + − cell means exceeding the + + raw score means by 2.49 to 4.15 units (approximately 3 months to over 1 year in grade-equivalent units) on the three subtests. Differences between the − + and − − treatment combinations are appreciably smaller for all strata. This second-order interaction illustrates that the significant Test Scorer × Previous Achievement Level interaction was generated almost exclusively in classes in which the teacher administered the test. It further indicates that Stratum 1 classes in which teachers administered the test preparatory to its being scored by outsiders produced pupil achievement scores notably above classes in the same stratum whose teachers both administered and scored the test themselves. At the other extreme were the observations in Strata 3 and 4. It might be suggested that teachers in Stratum 1

TABLE 4

RAW SCORES ON SUBTESTS OF STANFORD ARITHMETIC ACHIEVEMENT TEST: TEST ADMINISTRATOR × TEST SCORER × PREVIOUS ARITHMETIC ACHIEVEMENT (STRATUM) COMBINATION

Subtest	Previous arithmetic achievement (stratum)	Test Administrator × Test Scorer treatment combination								*F* ratio[a]
		++		+−		−+		−−		
		$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	
Computation	1 (High)	20.19	.74	22.68	3.85	21.38	3.93	21.69	1.18	1.57
	2	20.16	2.60	19.76	3.18	20.64	1.91	17.61	2.09	
	3	20.25	3.38	16.92	3.97	15.77	2.33	17.03	1.99	
	4 (Low)	14.93	2.09	13.43	4.09	12.53	1.18	13.01	1.91	
Concepts	1 (High)	17.31	1.01	20.98	2.39	18.01	2.39	18.51	1.44	3.08*
	2	16.85	1.47	16.90	2.12	17.44	1.46	16.16	.95	
	3	15.46	.60	12.93	.92	13.38	1.24	14.26	2.71	
	4 (Low)	12.04	.60	11.01	2.31	10.36	1.61	9.61	.94	
Applications	1 (High)	23.51	1.13	27.66	1.58	24.01	2.22	24.44	1.05	4.14*
	2	22.16	1.59	22.39	1.62	22.72	1.41	22.98	1.96	
	3	21.61	.83	17.59	1.15	18.59	2.09	19.47	3.57	
	4 (Low)	17.13	1.62	13.58	1.48	14.92	.83	13.08	1.69	
Average	1 (High)	20.34	.85	23.78	2.46	21.13	2.68	21.55	.62	3.43*
	2	19.72	1.84	19.69	2.29	20.27	.98	18.92	.90	
	3	19.11	1.41	15.81	1.99	15.92	1.80	16.92	2.34	
	4 (Low)	14.70	1.38	12.67	3.20	12.61	.32	11.90	1.39	

[a] Computed using initial error term (*df* = 16) for Computation Subtest; on all other subtests, final pooled error term (*df* = 31) was used.

* $p < .05$.

who administered the test may have taken procedural and motivational steps before and/or during their administration of the test to insure continued high achievement by their pupils even when the tests were scored by outsiders. On the other hand, teachers in the lower achieving strata may not have engaged in similar behaviors, or, if they did, these teacher behaviors had a minimal or even deleterious effect on the achievement scores of the pupils concerned.

It is important to keep in mind that the results reported above emerged from relatively simple manipulations of variables and were limited to standardized achievement tests in arithmetic. Aware of the differences that did emerge, investigators doing research in the schools should carefully standardize their procedures and techniques so as to control the effect of such methodological conditions.

REFERENCES

COOK, D. L. The Hawthorne effect in educational research. *Phi Delta Kappan*, 1962, 44, 116–122.

DOUGLASS, H. R. The effect of measurement upon instruction. *Journal of Educational Research*, 1935, 28, 508–511.

GOODWIN, W. L. The effects on achievement test results of varying conditions of experimental atmosphere, notice of testing, test administration, and test scoring. Unpublished doctoral dissertation, University of Wisconsin, 1965.

GREEN, B. F., & Tukey, J. Complex analysis of variance: General problems. *Psychometrika*, 1960, 25, 127–152.

HOPKINS, K. D., & LEFEVER, D. W. TV vs. teacher administration of standardized tests: Comparability of scores. Paper read at National Conference on Measurements in Education, Chicago, February 1965.

LOWELL, F. A preliminary report of some group tests of general intelligence. *Journal of Educational Psychology*, 1919, **10**, 323–344.

MAYO, E. *The political problems of an industrial civilization*. Boston: Harvard Business School, 1945. (a)

MAYO, E. *The social problems of an industrial civilization*. Boston: Harvard Business School, 1945. (b)

ORNE, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 1962, **17**, 776–783.

PHILLIPS, B. N., & WEATHERS, G. Analysis of errors made in scoring standardized tests. *Educational and Psychological Measurement*, 1958, **18**, 563–567.

RICE, J. M. The futility of the spelling grind. *Forum*, 1897, **23**, 163–172.

ROETHLISBERGER, F. J. *Management and morale*. Cambridge, Mass.: Harvard University Press, 1941.

ROETHLISBERGER, F. J., & DICKSON, W. J. *Management and the worker*. Cambridge, Mass.: Harvard University Press, 1941.

ROSENTHAL, R. On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. *American Scientist*, 1963, **51**, 268–283.

ROSENTHAL, R. Covert communication and tacit understanding in the E–S dyad. Paper read at American Psychological Association, Chicago, September 1965.

SIMPSON, R. H. The critical interpretation of test results in a school system. *Educational and Psychological Measurement*, 1947, **7**, 61–66.

THORNDIKE, R. L., & HAGEN, E. P. *Measurement and evaluation in psychology and education*. (2nd ed.) New York: Wiley, 1961.

TRAXLER, A. E. Administering and scoring the objective test. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D. C.: American Council on Education, 1951. Pp. 329–416.

(Received December 13, 1965)

Journal of Educational Psychology
1966, Vol. 57, No. 6, 359–365

MODIFIABILITY OF AN IMPULSIVE TEMPO[1]

JEROME KAGAN, LESLIE PEARSON, AND LOIS WELCH
Harvard University

Impulsive 1st-grade Ss were trained to be reflective under 2 different tutoring conditions. In 1 group of Ss (N = 20) the trainer first persuaded the child that he and the S shared some interests and attributes, while the 2nd group (N = 20) was not treated in any way that would lead the child to believe he and the trainer shared characteristics. Both groups showed longer response latencies after training. The condition of perceived high similarity to trainer facilitated the training for some girls, but not for boys.

An extended series of investigations on a conceptual tempo dimension called reflection-impulsivity has demonstrated impressive stability over both time and tasks, and meaningful predictions to a variety of problem-solving tasks (Kagan, 1965a, 1965b; Kagan, Rosman, Day, Albert, & Phillips, 1964). Impulsive children, in contrast to reflective children of similar age and verbal skills, make more errors in reading prose when in the primary grades, are more likely to offer incorrect solutions on inductive reasoning problems and visual discrimination tasks, and make more errors of commission on serial recall tasks. Impulsive children do not seem to care about making mistakes for they offer answers quickly and without sufficient consideration for the probable accuracy of their solution. This disposition is often a handicap in the typical school situation, for most teachers do not have a high tolerance for incorrect replies, and the peer group is prone to jeer at the child who impulsively blurts out obviously incorrect answers.

Although existing evidence suggests that an impulsive attitude begins its growth during the preschool years and may be a deeply entrenched habit (Kagan, Rosman, Day, Albert, & Phillips, 1964), it is worthwhile, nevertheless, to inquire into the modifiability of an impulsive tendency. This report summarizes one attempt to train impulsive children to be more reflective.

There are at least four major motives that mediate behavior change: (*a*) the desire for an external reinforcement, such as praise, love, money, toys, candy, etc., (*b*) the desire to avoid an unpleasant experience, (*c*) the desire to be correct and/or to be competent at a task, and (*d*) the desire to maximize similarity to a model.

The most popular incentives used by teachers, parents, employers, and psychologists to effect a change in response hierarchies capitalize on the first two motives, the child's desire for a positive social response and his desire to avoid a punishment. Some students of the tutorial process insist that the mere attainment of a correct answer is sufficient reward, and they remove all human agents from the learning situation.

The incentive least likely to be exploited is subject to controversy; namely, the child's desire to maximize similarity to a desirable adult. It is acknowledged that children spontaneously imitate and adopt selected responses displayed by particular adults. There is disagreement, however, as to the reasons for this imitation. It has been argued (Kagan, 1958)

---

[1] This research was supported in part by research grant MH-8792 from the National Institute of Mental Health, United States Public Health Service.

that the child wishes to maximize similarity to a desirable adult in order to share vicariously in his resources. This process eventually leads to an identification with the model. The motive to maximize similarity to a model appears to be of major importance in the establishment of unusually strong motivations for a career, especially intellectual careers. It was decided, therefore, to compare the differential effectiveness of training in reflection under two conditions, a normally nurturant condition between child and tutor, and one in which the child was persuaded to believe that he shared some attributes with the tutor and by becoming reflective he could increase the pool of shared characteristics. It was assumed that the initial belief in similarity to the tutor would act as an incentive to motivate the child to add to the number of shared similarities. If a child believed that the tutor valued reflection he should be more highly motivated to adopt this attitude than the child who perceived no initial similarities to the trainer.

This investigation asked two questions: Can an impulsive attitude be modified through direct training in reflection? Is there any training advantage when the child initially perceives some basis of similarity to the trainer?

METHOD

Procedure

Subjects. A group of 155 first-grade children was administered the Matching Familiar Figures (MFF) and two verbal scales from the Wechsler Intelligence Scale for Children, along with inductive reasoning tests, in the fall of the first grade. The MFF is a visual discrimination task in which pictures of familiar objects are presented along with six similar variants. The child must select the one variant that is identical with the standard. The child's response latency to the first solution hypothesis and errors are the major variables studied. Figure 1 illustrates a sample item from the test.

The subjects (*S*s) were classified as impulsive or reflective on the basis of their performance on the 12 MFF test items. Impulsive children were above the median on total errors and below the median on average response latency (for all 12 items) for their sex (that is, they responded quickly and made many errors). Reflective children were below the median on errors and above the median on response latency. The impulsive *S*s were assigned to one of three groups: trained under conditions of perceived similarity to the trainer (Group I-Id); trained under low perceived similarity to the trainer (Group I-non Id); and no training (Group I-C). The assignment of children to the three groups was such that the profile of verbal ability was close to identical across the groups. There were 10 boys and 10 girls in each of the three groups of impulsive children. In addition, 20 reflective children (10 boys and 10 girls) were selected from the total pool of 75 reflective children. These *S*s were not given any training and were treated as Group I-C.

Training

Each of the 40 children in the two training groups was seen by an adult the same sex as the child[2] and each trainer worked with children in both Groups I-Id and I-non Id. Each child was seen for three sessions, each session lasting about 40–50 minutes. The sessions were separated by 2 or 3 days.

Training session 1: Group I-Id. During the first part of the session the experimenter (*E*) began to persuade the child that *E* and *S* shared interests and attributes. This was accomplished by having *E* ask *S* a set of standard questions. After the child had answered *E* would answer in a similar way and comment on the similarity between adult and child. A sample of dialogue in Session 1 follows.

> How many brothers and sisters do you have? [After the child answers, *E* answers in a way consonant with the child and adds:]
>
> Gee, we are the same, we both have a (brother, sister). Let's see if we are the same in other ways. What are your favorite foods to eat? [After *S* answers, *E* responds in a consonant manner for one of the items and adds:]

---

[2] The trainers were Leslie Pearson, Lois Welch, David Cohen, and Robert Stewart.

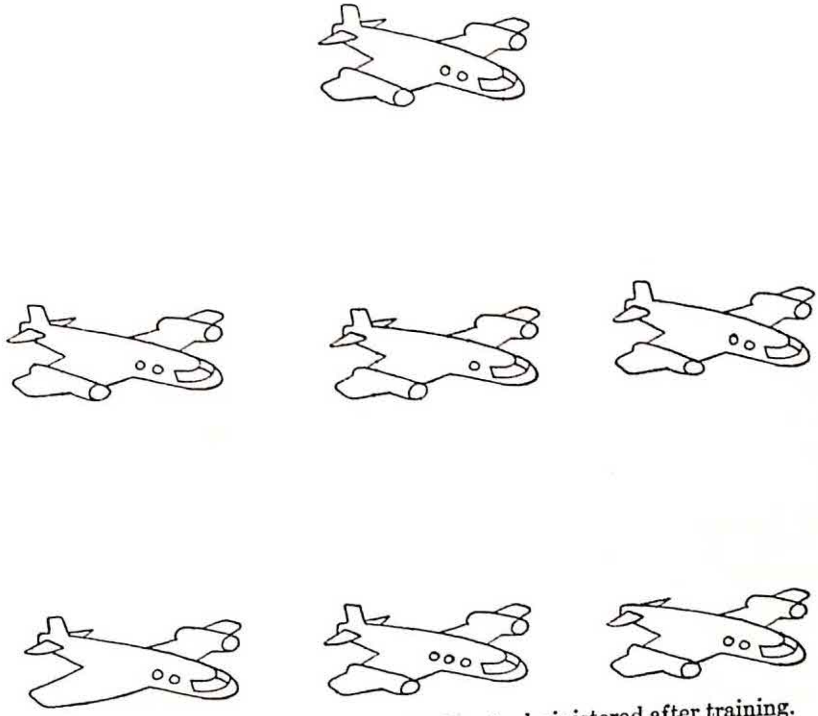


FIG. 1. Illustrative item from the MFF test administered after training.

> Oh, we're the same on foods, too. What's your favorite game? [After *S*'s answer, *E* answers in a consonant manner and remarks on the similarity and then asks:]
>
> What other games do you like to play? [*E* notes one consonance after these answers].
>
> Let's see how else we might be the same. What's your favorite animal? [After *S* answers, *E* answers consonantly, then brings out a leaflet of animal pictures with three pictures to a page].
>
> Let's look at some animal pictures that I have here and we will pick our favorites. [*E* then shows *S* three pictures, of a sea otter, squirrel, and giraffe. After the child picks a favorite animal *E* picks the same one and remarks on the similarity].

Following the first phase of similarity training, which took about 20–30 minutes, the specific training in reflection was initiated. The *E* first commented that he was reflective and valued reflection on tasks. His actual comments were:

> We're so much the same on so many things that I'm going to tell you something about myself. Maybe we can be the same on that, too. I think it's important not to make mistakes and to be right. So I take my time and always check my answers. One of the things we're going to do together is some other kinds of games a little bit like what you did just before vacation. The important thing in these games is to try not to make a mistake. I want you to get them right every time on your first try and one way to get them right is to think about your answer. [The *E* then demonstrated how he behaved reflectively on tasks with response uncertainty. The training then began].

The training tasks were the same in all three sessions and included parallel items. The three training tasks were a haptic-visual matching task, a design-matching task, and an inductive reasoning test.

In the design-matching task the child was shown a geometric design and an array of similar variants and was asked to select the variant that was exactly like the standard. This task had the same requirements as MFF except that all of the stimuli were geometric designs and none was a familiar figure. The second task was an inductive reasoning task in which the child was told three characteristics of an object and had to guess what the object was (e.g., What is cold, white, and round?). The third task was a haptic-visual matching task in which the child haptically explored a letter of the alphabet that was made from discrete tacks placed in a board. The child had to select from a visual display of letters the letter that was represented by the tack display he had explored haptically. The visual array was exposed while the *S* was exploring the tacks. During the training session *S* was given six items from each of these three tasks.

The training procedure for delay was direct. The *E* told the child that he was not allowed to offer an answer for a fixed period of time (10 or 15 seconds). The *E* said that he would look at his watch and tell the child when he was allowed to answer. The *E* instructed the child to study the stimuli in the task and to think about his answer during the enforced period of delay. The 15-second delay was imposed for all items on the visual matching and haptic-visual tasks; a 10-second delay was imposed for the inductive reasoning items. The child was only allowed one response per item and was not told if that response was correct or incorrect. All of the children delayed their responses most of the time. On the few occasions when a child did not delay, the *E* accepted the response and proceeded to the next item.

Training session 1: Group I-non Id. The *E* administered the same pretraining interview as described for Group I-Id, with one difference. The *E* never indicated that he shared interests or attributes with the child. He merely commented positively on the child's choice. The specific training in reflection proceeded exactly as was described for Group I-Id.

Training session 2. The format for training sessions 2 and 3 was similar to the one described for Session 1. The *S*s in groups I-Id and I-non Id were asked the same questions in the same order. However, the *S*s in Group I-Id were reminded how similar *E* and *S* were; the children in Group I-non Id were not given this information. The *E* spent the opening part of the interview asking new questions about *S*'s interests and, according to a fixed schedule, either commented that *S* and *E* were similar in their answers to selected items (for Group I-Id) or merely noted the child's answer without indicating consensual agreement. The direct training for reflection followed and the tasks included six additional items from the visual matching, haptic-visual matching, and inductive reasoning tests. As in Session 1, *E* insisted that the child delay a fixed period of time before answering.

Training session 3. The final session devoted a shorter time to pretraining matters. The training tasks included a final set of six items from the three training tasks.

In sum, the training procedure emphasized inhibition of impulsive answers, but placed no emphasis on more efficient visual scanning techniques or more analytic reasoning. The children were trained only to delay their answers; they were not given better ways of solving the problem.

Evaluation of Training

Each *S* was seen for a final evaluation session between 6 and 8 weeks after the third training interview. The examiners were two new adult females[3] who were complete strangers to the children. Moreover, the examiners did not know whether the children were impulsive or reflective nor to which training group they belonged. Each *S* was given an entirely new version of MFF under standard instructions and an inductive reasoning test. The items in the inductive reasoning test illustrated three pictures that described the beginning of a story or sequence. The child was also shown four additional pictures and was asked to select the one picture from the four alternatives that illustrated best "what happened next." Figure 2 illustrates a sample item from this test.

Data Analysis

Each *S* had been given different forms of the MFF and Picture Completion Reasoning Tests six months earlier, in the fall of the first grade. The major variables analyzed were: (*a*) the raw response time and error scores following training, and (*b*) change scores for response time and errors pre- and posttraining. The MFF variables included absolute and percent-change scores for response time and errors (percent-change score was obtained by dividing the difference between the values on the first test session and the second test session by the first test session score). For the Picture Completion Reasoning Test the variables were mean number of errors made after training, response time to the incorrect choices, response time to all items, and the percent-change score for response time to the incorrect items.

---

[3] Janet Levine and Marion Cleveland were the examiners for this evaluation phase.

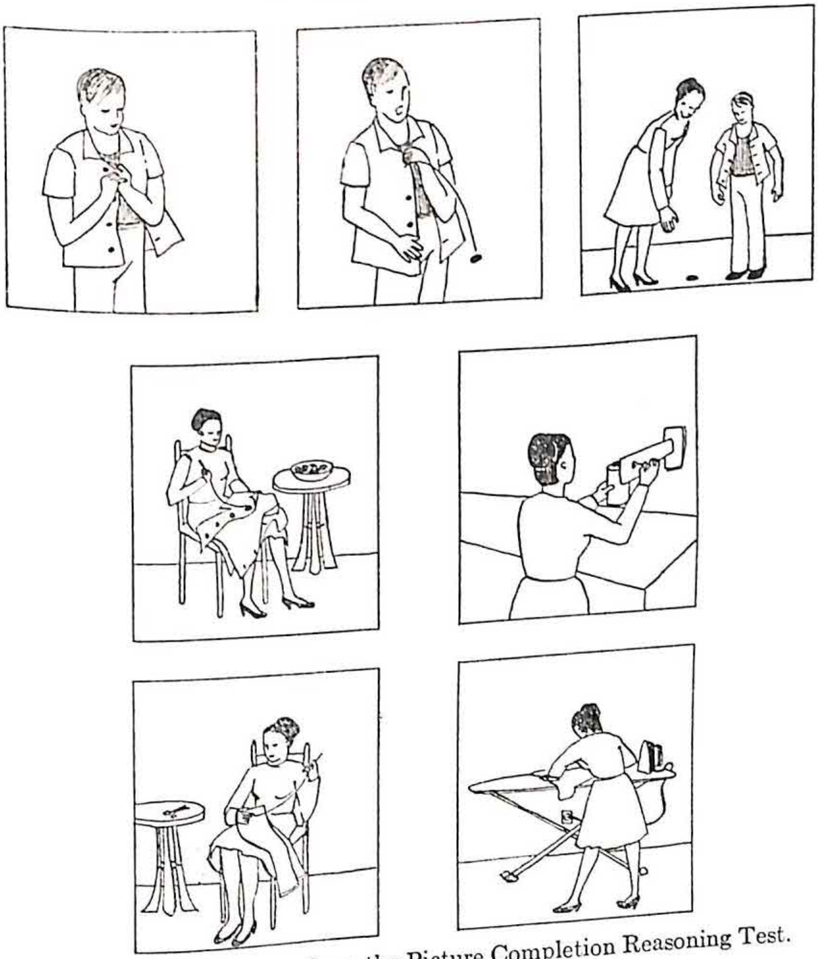


FIG. 2. Illustrative item from the Picture Completion Reasoning Test.

RESULTS AND DISCUSSION

Table 1 contains the mean scores for each of the groups on the major variables. *The only important effect of training was to lengthen the response times to MFF.* The children in each training group showed longer response times to MFF than the impulsive controls ($p < .05$, two-tailed test for each group). Moreover, the trained group showed larger percent increases in response time, when the current response time measure was compared with the one obtained 6 months earlier ($p < .05$, t test for Group I-Id; $p < .10$ for Group I-non Id). It is of interest that the response times for the reflective children who were not seen between fall and spring were not significantly longer than those of the impulsive children who were trained in delay. The training produced response latencies that matched those of normal reflectives.

The variability of the MFF response time scores was much greater after training than it was in the fall, but this increase was due, in part, to the large increase in mean response times for Group I-Id and Group I-non Id. The coefficient of variation ($100\ SD/M$) on the early fall test was 30; the average coefficient of variation after training was 63 for Group I-Id, 64 for Group I-non Id, and 53 for the control group (Group C).

TABLE 1
MEAN SCORES ON MFF AND PICTURE COMPLETION REASONING TESTS

Variables	Group I-Id		Group I-non Id		Group I-C		Group-Reflective	
	Boy	Girl	Boy	Girl	Boy	Girl	Boy	Girl
MFF								
Response time								
M	15.2	32.7	13.2	21.6	7.2	14.0	13.8	28.8
SD	7.8	21.2	8.8	13.8	3.4	8.0	7.3	16.4
Errors								
M	28.0	16.6	25.1	18.3	27.6	23.9	22.6	17.8
SD	7.6	4.7	7.1	5.5	7.6	8.8	5.4	5.6
% increase response time	133.4	460.0	127.1	236.6	36.2	137.3	19.3	65.1
% increase errors	5.7	−25.8	4.0	−19.3	22.0	−4.5	165.7	102.1
Picture completion								
Errors								
M	8.0	4.8	7.4	5.5	8.4	5.3	5.0	4.0
SD	2.8	2.8	3.5	2.9	2.5	2.8	2.6	3.4
Response time to incorrect								
M	11.7	17.9	15.1	18.9	12.2	16.5	15.8	16.3
SD	4.6	8.5	5.9	6.6	4.8	5.0	5.7	12.4
Response time, all items								
M	10.9	14.9	13.5	16.0	11.8	15.0	12.8	15.5
SD	3.4	3.6	3.7	4.0	4.6	4.0	4.7	3.8
% increase in response time to incorrect	37.4	70.7	70.3	110.2	0.1	86.2	27.3	−2.8

Although both the absolute and relative variability of response times were greater in the spring, there was no marked difference in variability among the three groups of impulsive children.

Unfortunately, error scores were not much affected by the training. The results were in the anticipated direction, for both training groups made fewer errors than the nontrained controls, but the differences did not reach acceptable levels of significance. Percent-change in errors revealed that the girls showed an average decrease in errors, whereas the boys' average change score was low, but positive. In sum, training had a marked effect on response delay, but a minimal effect on quality of performance. Since the primary focus of the training was on inhibition of a fast response rather than improved strategies of visual detection, it is reasonable that response latency should be affected more dramatically than recognition errors.

Effect of Training on Inductive Reasoning

The *S*s were not given any training specific to the items in the Picture Completion Reasoning Test, but it was hoped that the training would affect performance on this task. As with MFF, the training produced increased response times to incorrect choices, but the differences were not statistically significant. Errors were not affected by the training.

Effect of Perceived Similarity to the Trainer

The differential effect of training under high or low perceived similarity was of borderline significance. It will be recalled that the children were trained by an adult the same sex as themselves, but it was only possible to use female adults as the assessors in the final evaluation. This condition of testing should lead, a priori, to an attenuation of the effect among boys and a maximization of the effect for girls, since the girls were trained and tested by an adult of the same sex. The results are congruent with this expectation, for the effects of high or low perceived similarity during training were more dramatic for girls than for boys on both tests. The mean percent increase in response time on MFF was 460% for Group I-Id girls, in contrast to 237% for Group I-non Id girls. Although these means were not significantly different, the girls at either extreme supported the major hypothesis. Of the total of 20 impulsive girls in Groups I-Id and I-non Id, 4 of the 5 who manifested the greatest increase in response time after training were in Group I-Id; whereas 4 of the 5 who decreased the most in response time were in Group I-non Id. This 4–1–1–4 distribution yielded a one-tailed p value of .103 by the Exact Test. The magnitudes of response time increases shown by these four girls in Group I-Id also furnish some support for the hypothesis. The actual change scores in Group I-Id were as follows: from 4 to 80 seconds, from 5 to 58 seconds, from 6 to 31 seconds, and from 6 to 29 seconds.

In sum, the brief training (about 60 minutes total) in delay produced larger response latencies among impulsive children. Unfortunately, the training did not have a strong effect on error scores and did not generalize to the inductive reasoning test. Since the adult who administered the posttraining MFF was not the trainer, the results cannot be attributed to a desire to please the trainer. The tendency to delay did generalize to another adult. The study demonstrates that impulsive children can be taught to modify their behavior and the effect can last a few weeks. The hypothesis concerning the facilitating effect of perceived similarity did not receive convincing support. The effect was manifest for only a few of the girls. However, the authors believe that these minimal trends warrant continued study of the psychological significance of perceived similarity between tutor and tutee.

REFERENCES

KAGAN, J. The concept of identification. *Psychological Review*, 1958, 65, 296–305.

KAGAN, J., ROSMAN, B. L., DAY, D., ALBERT, J., & PHILLIPS, W. Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs*, 1964, 78(1, Whole No. 578.)

KAGAN, J. Impulsive and reflective children. In J. D. Krumboltz (Ed.), *Learning and the educational process*. Chicago: Rand McNally, 1965, 133–161. (a)

KAGAN, J. Reflection-impulsivity and reading ability in primary grade children. *Child Development*, 1965, 36, 609–628. (b)

(Received January 5, 1966)

Journal of Educational Psychology
1966, Vol. 57, No. 6, 366–372

RELATIONS AMONG PHYSICAL SKILLS AND PERSONAL AND SOCIAL VARIABLES IN THREE CULTURES OF ADOLESCENT GIRLS[1]

BOYD R. McCANDLESS
Emory University

AND

FAHMIDA ALI
Lady Maclagan College, Lahore, West Pakistan

To test the hypothesis that physical skills have different value for girls in sex-segregated schools from different cultures than they have in coeducational schools, buddy ratings of physical skills and social acceptance, school grades, standardized achievement test, anxiety, and freedom from lying scores were obtained from 3 samples of 9th-grade girls, 2 in religious sex-segregated schools (Muslim Pakistani, N = 60; U.S. Roman Catholic, N = 60; and U.S. coeducational mixed religions, N = 52). Academic aptitude test scores were available for most of the 2 United States samples. 10 of the 14 possible correlations were homogeneous. Social acceptance, academic achievement, school marks, and academic aptitude correlated positively with each other. The 4 correlations were nonhomogeneous: academic aptitude—social acceptance; academic aptitude—physical skills; social acceptance—school grades; and social acceptance—physical skills.

The present study was designed to test the hypothesis that physical skills have different values for girls in sex-segregated schools from different cultures than they have in coeducational schools.

The authors have been able to find few data bearing directly on this problem, either for United States adolescent girls or for girls of any other nationality. However, the consensus of the sparse research for adolescent girls suggests that proficiency in sports and games has, if any, only low positive relationships within groups of girls to such factors as their intelligence, their academic achievement, or their popularity.

Kuhlen and Lee (1943) present data showing percentages of high-popular and low-popular girls possessing certain physical skills, attributed to them by "Buddy Ratings." Bretsch (1952) gives the percentage distributions of high and low popularity quartile ninth-grade girls who rated themselves on participation in and/or skill in tennis. This is the only physical skill he reports that (impressionistically) seems similar to badminton, one of the skills used for Pakistani girls in the present sample. All subjects (*Ss*) in the present study judged *others* in their class by the buddy-rating technique, but did not report on themselves. Hence, the results reported by Bretsch are not comparable to the present sample. However, those of his girls ($N = 371$) reporting themselves as participating in tennis were significantly more popular than those who reported that they did not play tennis. This may be a function of the 'general extracurricular syndrome = popularity syndrome', or to the social-class-relatedness of playing tennis (i.e., tennis is a relatively upper-class sport, and popularity is positively related to social-class status).

---

[1] The authors appreciate the cooperation extended to them by the administrations, staffs, and students of Lady Maclagan College, Lahore, West Pakistan, and Our Lady of Grace and University Schools, Indianapolis and Bloomington, Indiana, respectively.

Kuhlen and Lee, using the category, "Active in Games," as rated by buddies for ninth-grade girls, find *S*s high in popularity (57%) "Active in Games," while those low in popularity are not (17%). This suggests a low positive relation, but the methods are so different from those used in the present study that no meaningful comparison can be made.

Somewhat more direct evidence exists for a positive relation between intelligence and social acceptability. Thorpe (1955) and Damrin (1949), respectively, report for samples of British children averaging about a year younger than the present sample, and of American girls who average about a year older. A number of subsample correlations are reported, which for the British sample range from −.35 to +.53, with an estimated overall correlation of .15; and for the U.S. (Damrin) sample .31. However, Damrin's sample ranged in age from Grades 9 through 12, the median age being about 16, while the present sample is homogeneous in grade level.

METHOD

Subjects

In the present study, the procedure for gathering data from the three different samples was as similar as possible. Details are given for the methods of collecting the first set of data. Differences in procedure are mentioned when they occurred for the second and third sets of data. The first sample (Un) was composed of middle- and upper-middle-class ninth-grade girls in a coeducational public school in a Midwestern university town having a population of about 40,000. Most of the girls were Protestant. The second set of data (Ca) was gathered from urban middle- and upper-middle-class ninth-grade girls enrolled in an urban Catholic sex-segregated private school, where most of the students were day students. These girls come from a city culturally very similar to that of the Un girls. The two towns are about 50 miles apart, and much intercommunication exists between them. The third sample was made up of middle- and upper-middle-class ninth-grade urban Muslim Pakistani (Pa) girls in a sex-segregated school supported by a combination of private and public funds. The Pa sample consisted entirely of day students and came, of course, from a culture extremely different from the Un and Ca samples.

Procedure

All girls in all samples had known each other for at least that portion of the academic year preceding the gathering of data (about 8 months). Data for the Un and Pa girls were gathered by the women physical education teachers in the respective schools (the junior author for the Pa sample). The Sister-Principal and the senior author worked cooperatively in gathering data for the Ca sample. Almost all girls were 14 or 15 years old, and were estimated to range from average to superior in intelligence. The Results section gives intelligence test data for those girls for whom they were available.

Not all girls took all tests, so the data are reported for a maximum number of 52 girls for the Un sample and 60 girls each for the Ca and Pa samples. Where relevant, numbers are indicated in all cases in the tables or text.

The following data were gathered, confidentiality being assured and facilities provided for the girls to work in seclusion:

1. Popularity: Each girl rated every other girl in her class from 1 (one of my best friends) to 5 (not my friend at all). The most popular girl in the Ca and Pa samples could thus achieve a score of 59 (if every other girl ranked her 1), and the least popular 295. For the Un sample, these figures are 51 and 255.

2. Academic achievement: Average school marks in the "solid" school subjects for the first semester of the ninth-grade academic year were used as the criterion for the Un and Ca samples (a grade of A was weighted 4.0; F, .0). Ninth-class middle-year marks, based on a Government devised, nationally uniform, "external" examination over the solid subjects were used to determine academic achievement for the Pa sample.

3. Physical skills: Skill at volleyball, tennis, broad jump, and sprint formed the buddy-ratings used for the Un and Ca samples. Net ball (similar to volleyball) and badminton (tennis was not available to the Pa sample, although the game is widely played in Pakistan by men and boys) were used rather than volleyball and tennis for Pa

sample. Each of these activities was rated by her "buddies" from 1 (a "best player") through 4 ("below-average player") and 5 ("never plays at this game"). The final proficiency score was the sum of these ratings for the four activities. For the Pa sample, phi between the physical education teacher and the average buddy-rating was .70. Buddy-ratings are used as the measure of physical skills for each sample.

For 24 of the Un girls (but for none of the other samples), actual measures (of unknown reliability) existed for proficiency in dash, broad jump, and "composite physical skills." Buddy-ratings correlated .69 for dash versus performance, .73 for broad jump versus actual performance; and .61 for "composite physical skills" versus performance. None of these correlations is corrected for attenuation, although each is significant below the .01 level.

4. Anxiety and Lie (here referred to as Freedom from Lying and probably actually "lack of defensiveness") scores were obtained for the Un and Ca samples, using the Children's Manifest Anxiety Scale (Castaneda, McCandless, & Palermo, 1956). Translation into Urdu was made for the Pa sample. In this process, one of the 11 Lie items was lost, since two of the items translate identically into Urdu, although they differ in English.

5. Scholastic Aptitude: Commonly used, standardized tests of academic aptitude (intelligence) were available for a majority of the Un and Ca samples, more complete coverage existing for the Un girls. A semi-standardized test modeled on existing United States tests but translated into Urdu was administered to most of the Pa sample. All tests on which results are reported had been administered late in the eighth or early in the ninth grade for the Un and Ca samples, and late in the ninth grade for the Pa girls.

6. Standard tests of academic achievement: These were available for a majority of the Un sample, about half of the Ca sample, and none of the Pa sample.

The entire set of data was collected for the Ca sample in 1 day, and within an academic week for the other two samples.

RESULTS

The results of interest in this study lie principally not in norms, but in relations among the different variables. Table 1 presents intercorrelations separately for the Un, Ca, and Pa populations.

Of the 14 first-order correlations run for all three samples, 10 were homogeneous (Snedecor, 1946), indicating substantial similarity among the groups. Homogeneity, to begin the listing with the Academic Aptitude (intelligence) variable, existed for its relations with school grades (r's ranging from .58 for the Pa sample to .71 for the Ca sample, each r being significant at the .01 level or less). Its demonstrated relations with the Anxiety measure were also homogeneous, r's ranging from $-.15$ for the Ca sample to $-.32$ for the Pa sample. These negative r's were statistically significant at the .05 level or less only for the Pa sample.

Intelligence and Freedom from Lying also showed homogeneous r's, ranging from .08 for the Ca girls to .38 for the Un sample, significance at .05 level or less existing only for the last-named sample.

The sociometric variable (Social Acceptance) was homogeneous and essentially zero in its relations with Anxiety. Social Acceptance was also homogeneous in its correlations with Freedom from Lying, positively related for each sample, but significantly so (at the .01 level or less) only for the Pa sample, where it reached .43.

School Grades followed the pattern for Academic Aptitude: Homogeneous but nonsignificant r's existed between composite marks and Physical Skills for all three samples (ranging from .01 for the Un sample to .23 for the Pa group); and Anxiety (the range of 4's was from essentially zero for the Un and Pa girls to a statistically significant $-.36$ for the Ca group). School Grades and Freedom from Lying followed much the same pattern: essentially zero for the Ca

TABLE 1

CORRELATIONS FOR THE UNIVERSITY SCHOOL, CATHOLIC, AND PAKISTANI SAMPLES AMONG THE DIFFERENT VARIABLES

	Social acceptance	Achievement test	School grades	Physical skills	Anxiety	Freedom from lying
Academic aptitude						
University	$N = 51$ $.39^{**}$	$N = 46$ $.76^{**}$	$N = 50$ $.63^{**}$	$N = 51$ .16	$N = 51$ −.25	$N = 51$ $.38^{**}$
Catholic	$N = 42$ .04	$N = 30$ $.62^{**}$	$N = 42$ $.71^{**}$	$N = 42$ −.06	$N = 37$ −.15	$N = 37$ .08
Pakistani	$N = 45$ $.56^{**}$		$N = 45$ $.58^{**}$	$N = 45$ $.41^{**}$	$N = 45$ $-.32^{*}$	$N = 45$ .23
Social acceptance						
University		$N = 47$ .11	$N = 51$ $.40^{**}$	$N = 52$ $.28^{*}$	$N = 52$ −.02	$N = 52$ .10
Catholic		$N = 30$ .32	$N = 60$ $.30^{*}$	$N = 60$ $.66^{**}$	$N = 55$ −.19	$N = 55$ .17
Pakistani			$N = 60$ $.64^{**}$	$N = 60$ $.58^{**}$	$N = 60$ .09	$N = 60$ $.43^{**}$
Achievement test						
University	$N = 47$ −.15		$N = 47$ $.78^{**}$	$N = 47$ −.01	$N = 47$ −.11	$N = 47$ $.31^{*}$
Catholic	$N = 30$ −.32		$N = 30$ $.79^{**}$	$N = 30$ .15	$N = 28$ −.08	$N = 28$ .13
Pakistani						
School grades						
University				$N = 51$ .01	$N = 51$ .02	$N = 51$ .17
Catholic				$N = 60$ .20	$N = 55$ $-.36^{**}$	$N = 55$ −.02
Pakistani				$N = 60$ .23	$N = 60$ .04	$N = 60$ $.35^{**}$
Physical skills						
University					$N = 52$ .04	$N = 52$ .02
Catholic					$N = 55$ .21	$N = 55$ .07
Pakistani					$N = 60$ .00	$N = 60$ $.32^{*}$
Anxiety						
University						$N = 52$ $.29^{*}$
Catholic						$N = 55$ $.35^{**}$
Pakistani						$N = 60$ .04

$^{*}p < .05$.
$^{**}p < .01$.

and Un samples, and a statistically significant ($p < .01$) .35 for the Pa group.

Relations between Physical Skills and Anxiety were homogeneous among the three samples, and in no case did the relation differ significantly from zero. The same relations held for Freedom from Lying and Physical Skills, except that the correlation for the Pa sample reached a .05 level of significance of .32.

Standardized academic achievement tests were available only for the Un and Ca samples. These relations were homogeneous for all variables: Academic Aptitude, positive and significant at below the .01 level for both samples: Social Acceptance, nonsignificant for both samples; School Grades, positive and significant at below the .01 level for both samples—.78 and .79 respectively; Physical Skills, nonsignificant for both samples; Anxiety, nonsignificant for both samples; and Freedom from Lying, positive and significant at the .05 level for the Un sample, nonsignificant but positive for the Ca sample.

The four sets of relations found to be nonhomogeneous are of particular interest to the authors. Two of these involved Academic Aptitude ("intelligence") and other variables, and the other two involved Social Acceptance. These nonhomogeneous correlations should reveal cross-cultural differences, if such exist in accordance with the authors' speculations.

The first, Academic Aptitude and Social Acceptance (popularity) lends itself poorly to plausible speculation (relations were positive and significant at below the .01 level for the Un and Pa samples, but approached zero for the Ca sample). It might be suggested that the Un and the very highly selected Pa samples value brains in the social and possibly marital market place, but that Academic Aptitude plays a less important role for well-protected, middle- and upper-middle-class, sex-segregated Catholic girls who do not, as the Un girls do, have to compete directly with boys or, as in the case of the Pa girls, to prove themselves (as a true minority group) equal to or better than boys.

The second, Academic Aptitude and Physical Skills, approaches zero for the Un and Ca samples but is positive and significant at less than the .01 level for the Pa sample. This suggests that the Pakistani girl is trying to demonstrate that she can hold her place among males.

The third set of relations, between Social Acceptance and School Grades, was positive and significant below the .05 level for all three samples. However, the Ca sample showed a significantly lower correlation than the Pa sample (.30 and .64 respectively, significantly different at the .02 level), and the Un sample and Pa sample where the difference closely approached significance at the .05 level (r's were .40 and .64 respectively). This pattern of relations is plausible: To a greater degree for the Pakistani than for the American girl, education means status; success in school represents emancipation and, perhaps, successful competition with the male. Hence, it is more valued as a criterion for social acceptability than for United States girls. Maternal modeling may also enter the picture, as there is a strong trend, regardless of fathers' social class, for only those Pa girls whose mothers are educated to reach ninth grade themselves.

Fourth, Social Acceptance is highly and positively correlated with Physical Skills for both sex-segregated samples (.66 for the Ca and .58 for the Pa sample), but barely reaches significance at the .05 level ($r = .28$) for the coeducational (Un) sample. The difference between Un and Ca girls is significant at the .01 level, and almost reaches .05 significance for the Un and Pa samples. In a girls' world, does athletic ability (a "masculine" trait) lend status, while it is of less importance in a coeducational world where the competition against others of the same sex may be greater for the attention of boys than in the area of physical skills?

In no case were these correlation patterns altered when partial correlations were run, indicating that the pattern of relations is not a function of a halo effect.

DISCUSSION

Striking sample-intercorrelation similarities exist among the three samples of U.S. University School ninth-grade girls (Un), Roman Catholic parochial school, sex-segregated ninth-grade girls (Ca), and Muslim sex-segregated ninth-grade Pakistani school girls (Pa). This suggests that translation of research instruments from one to another language (Urdu for the Pa sample), and their use for populations from very different school settings may be quite feasible. For example, a measure of Academic Aptitude (intelligence) and school marks in the so-called solid subjects showed homogeneous correlations among the three groups, uniformly high and positive (Un = .63, Ca = .71, and Pa = .58). From a total of 14 correlations between different variables calculated for all three cultures (Academic Aptitude, Social Acceptance, School Grades in the solid subjects, an index of Physical Skills, Anxiety, and Freedom from Lying—'Lack of Defensiveness'), only four sets of correlations were nonhomogeneous.

For the two United States samples, standardized tests of academic achievement were also available. Correlations between scores and all other variables mentioned in the preceding paragraph were homogeneous for the Un and the Ca samples, ranging from high positive for Academic Achievement Tests versus School Grades (approaching $r = .80$) to almost zero (Academic Achievement Test versus Sociometric Status, Physical Skills, Anxiety, and Freedom from Lying).

These results are pleasantly surprising from a research worker's point of view when one considers the great cultural range. The Pa sample came, for example, from a Muslim culture, predominantly "purdah"—that is, veiled, exposed after childhood only to males from the closest family circle; was administered Government composed examinations, with no teacher influence on grades; and spoke English and/or were exposed to Western norms little or not at all. Residence was in a great city of more than one million population. In contrast, the Un sample was enrolled in an experimental, demonstration coeducational school in a small Midwestern, university-centered town; and the Ca sample was sex-segregated and resided in a moderately urban area (about one-half million residents).

The nonhomogeneous correlations are of particular interest to the authors. They numbered four:

1. Academic Aptitude and Social Acceptance, significantly positive for the Un and Pa samples, essentially zero for the Ca sample. This suggests that the coeducational University School community imbues intelligence with social desirability, the Pakistani-Muslim community regards it as a means whereby "women

may claim their own," and the Catholic girls' community views it relatively neutrally.

2. Academic Aptitude and Physical Skills correlated about zero for the United States samples, but significantly positively for the Pakistani sample. The brighter, relatively suppressed Pakistani girl may believe that, to succeed, she must excel in everything and, in her deeply sex-segregated culture, be in competition with the male, fighting him for status.

3. The Pakistani girls may also be struggling, along with the coeducational girls, for intellectual status, and appreciating this among each other, where the Catholic girls are less concerned with such striving.

4. In a unisexual culture (the Catholic and Pakistani girls), physical prowess may provide a source of group status that is of little importance to the coeducationally reared girl, who may value the attention of boys more than her success in the motor area.

REFERENCES

BRETSCH, H. S. Social skills and activities of socially accepted and unaccepted adolescents. *Journal of Educational Psychology*, 1952, **43**, 449–458.

DAMRIN, D. E. Family size and sibling age, sex, and position as related to certain aspects of adjustment. *Journal of Social Psychology*, 1949, **29**, 93–102.

KUHLEN, R. G., & LEE, B. J. Personality characteristics and social acceptability in adolescence. *Journal of Educational Psychology*, 1943, **34**, 321–340.

SNEDECOR, G. W. *Statistical methods* (4th Ed.). Ames, Iowa: State College Press, 1946.

THORPE, J. G. An investigation into some correlates of sociometric status within school classes. *Sociometry*, 1955, **18**, 49–61.

(Received January 14, 1966)

Journal of Educational Psychology
1966, Vol. 57, No. 6, 373-381

TEACHERS' BELIEF SYSTEMS AND PRESCHOOL ATMOSPHERES[1]

O. J. HARVEY
University of Colorado

B. JACK WHITE
University of Utah

MISHA S. PRATHER
University of Colorado

RICHARD D. ALTER
University of Utah

AND JAMES K. HOFFMEISTER
University of Colorado

With the aim of determining the effects of teachers' belief systems upon the classroom atmospheres they create for preschool children, 30 teachers, 10 in each of 3 levels of concreteness-abstractness, were observed and rated by pairs of independent judges on 26 dimensions relating to such factors as flexibility, attitudes toward rules, encouragement of the children's independence and creativity, need for structure, and punitiveness. The more abstract teachers differed from the more concrete Ss in what is presumed to be an educationally favorable direction on all dimensions.

Teachers differing in the concreteness-abstractness of their belief or personality systems should also differ in the goals, social milieu, and behavior they seek for and from their students and hence, knowingly or unknowingly, in the values and content they communicate, in their styles of communication and in their reactions to student adherence to or departure from the standards embodied in their beliefs. Because concreteness disposes towards more fixed and categorical belief systems (Harvey, 1966; White & Harvey, 1965), toward an authority orientation rather than a task orientation (Harvey, 1966; Tiemann, 1965), and toward seeking a simple and highly structured environment (Harvey, 1965; Ware & Harvey, in press), more concrete teachers, to a greater extent than more abstract teachers, should impose predetermined goals upon the students, provide structure and detailed means for their attainment, be less tolerant of student deviation from their goals and standards, and, consequently, react more strongly and invariantly to such deviation. A test of some of the more specific aspects of this general hypothesis was the purpose of this study.

METHOD

Head Start teachers differing in concreteness-abstractness according to two different measures were rated on 26 dimensions assumed to reflect educationally desirable and undesirable behavior toward their preschool students. The rating categories were (1) expression of warmth toward the children, (2) perceptiveness of the children's wishes and needs, (3) flexibility in meeting the needs and interest of the children, (4) ability to maintain relaxed relationships with the children, (5) attention to the individual child, (6) task involvement, (7) enjoyment of teaching, (8) enlistment of child participation, (9) encouragement of individual responsibility, (10) encouragement of free expression of feelings, (11) encouragement of creativity, (12) teaching new concepts, (13) ingenuity in improvising teaching and play materials, (14) utilization of physical resources, (15) task effectiveness, (16) diversity of activities simultaneously permitted, (17) smoothness of classroom operation (especially in the transition from one activity to another), (18) consistency of rule enforcement, (19) use of functional explanation of rules, (20) use of nonfunctional explanation of rules, (21) use of unexplained rules, (22) rule orientation, (23) determination of classroom and playground procedure, (24) need for structure in teaching activities and relationships with children, (25) punitiveness, and (26) anxiety induced by the observers' presence.

---

[1] This study was supported by the Head Start Program, Office of Economic Opportunity, under Contract OEO-517 with the University of Colorado Extension Division. This study in whole or part may be used for any purposes of the United States Government. We wish to express our thanks to OEO and Project Head Start for their support, to the many observers for their conscientious effort in collecting the data and to the teachers in and directors of Head Start Programs that permitted us to visit their schools and observe their teaching.

It was anticipated that the more abstract subjects (*S*s) would score higher than the more concrete *S*s on dimensions 1–19 and lower than more concrete *S*s on dimensions 20–26.

Subject Selection

Teachers were selected for subsequent observation on the basis of their responses to the "This I Believe" (TIB) Test scored according to conceptual systems.

The TIB, developed specifically as a measure of conceptual or belief systems (e.g., Harvey, 1964, 1965, 1966; White & Harvey, 1965), requests *S* to indicate his beliefs about a number of socially and personally significant concept referents by completing in two or three sentences the phrase, "This I believe about ________," the blank being replaced successively by one of the referents. The referents employed in the present study were comprised of "religion," "friendship," "the American way of life," "sin," "education," "the family," "people on welfare," "punishment," "teaching," and "sex"; the last six substituted for the more standard referents with the aim of insuring their specific relevance to the *S*s and to the task of teaching in the Head Start Program.

From the relativism, tautologicalness, novelty, and connotative implications or richness of the completions, together with criteria implied below, respondents may be classified into one of the four principal belief systems posited by Harvey, Hunt, and Schroder (1961) or into some admixture of two or more systems.

More specifically, *S*s were classified as representing *predominantly* System 1, the most concrete mode of dimensionalizing and construing the world, if their sentence completions denote such characteristics as high absolutism, high tautologicalness, high frequency of platitudes and normative statements, high ethnocentrism, high religiosity, assertion of the superiority of American morality, and expression of highly positive attitudes toward institutional referents.

The *S*s were categorized as representing System 2, the next to the lowest level of abstractness if, in addition to being highly evaluative and absolutistic, they expressed strong *negative* attitudes toward such referents as marriage, religion, and the American way of life, the same referents toward which System 1 representatives manifest highly *positive* attitudes.

Responses to the TIB were scored as representing System 3 functioning, the next to the highest level of abstractness posited by Harvey et al. (1961), if they indicated more relativism and less evaluativeness than Systems 1 and 2 and at the same time expressed strongly positive beliefs about friendship, people, and interpersonal relations.

System 4 functioning, the highest of the four levels of abstractness, was indicated by TIB responses that implied a high degree of novelty *and* appropriateness, independence without negativism, high relativism and contingency of thought, and the general usage of multidimensional rather than unidimensional interpretive categories.

The TIB was administered to the 168 teachers participating in the Head Start training program conducted by the University of Colorado Extension Division during the summer of 1965. From among this number, no instances of System 2, the anti-authority orientation, and only 10 cases of System 4 functioning were found. Ten *S*s each from System 1 and 3, selected from a somewhat larger subpopulation of representatives of these systems, were added to the 10 System 4 representatives, for a total of 30 experimental *S*s, all women.

An Additional Measure of Concreteness-Abstractness

In addition to the TIB, *S*s completed an objective measure of concreteness-abstractness, the Conceptual Systems Test (CST), which was developed through Tryon's cluster and factor analysis (Tryon & Bailey, 1965[2]). Four factors were found which are theoretically consistent with the major characteristics of the four principal conceptual systems or levels of concreteness-abstractness posited by Harvey et al. (1961). These factors, as we have tentatively labeled them, together with some of their representative items, are:

---

[2] R. C. Tryon and D. E. Bailey, "The B. C. Try system of cluster and factor analysis: Multivariate behavioral research," in preparation.

1. *Divine fate control or religious fundamentalism.* This is assessed by such items as "There are some things which God will never permit man to know," "In the final analysis, events in the world will be in line with the master plan of God," and "I believe that to attain my goals it is only necessary for me to live as God would have me live."

2. *Need for certainty or simplicity* is expressed in response to such statements as "I prefer a story that has two themes rather than one that has five or six themes going at once," "People who seem unsure and uncertain about things make me feel uncomfortable," and "The effective person is one who does not hold conflicting beliefs."

3. *Tolerance of complexity and uncertainty* is based on such items as "I have so much trouble finding out what is or is not true that I can't understand how some people can feel so certain that they know the truth," "More often than not, I like some aspects of a person and do not like other aspects of him," and "I find that I cannot help analyzing almost everything I see and hear."

4. *Relativism of truth* is measured by such items as "There can be as many truths as there are individual points of view," "Man is the judge of the truth or untruth of his thoughts and behavior," and "Something is true or untrue depending on one's assumptions and the context."

Of the several ways these factors may be scored, the one utilized for this report was the overall abstractness score, represented by the sum of the item responses on a six-point scale (from "Completely disagree" to "Completely agree") across the four factors.

Subjects

Ten representatives each of System 1, 3, and 4, as selected by the TIB, served as the experimental *S*s. The mean score of each system on the CST, the second measure of abstractness, was: System 1, 2.98; System 3, 3.43; and System 4, 4.35. According to *t* tests, each system differed significantly from each of the others ($p < .05$ for all comparisons).

All *S*s had had prior teaching experience, System 1 teachers having taught on the average almost twice as long (10.1 years) as teachers from System 3 (5.8 years) or System 4 (5.4 years). However, only one System 1, two System 3 and three System 4 teachers had taught previously at the preschool level. All *S*s participated in a common 1-week training program for Head Start teachers shortly before beginning their teaching of Head Start enrollees.

The geographic locations where the *S*s were teaching and were observed ranged from urban housing projects to small towns which served predominantly rural children in their Head Start programs. In an attempt to control for the influence of different administrative structures and physical facilities upon the teachers' classroom behavior, an equal number of teachers of each belief system was selected from a common administration and a common building whenever possible. In the three instances in which representatives of the three belief systems could not be obtained from the same program, a teacher was selected from an alternate program that served children comparable in socioeconomic backgrounds to the children being taught by teachers of the other two belief systems.

The Rating Scale

Teachers were rated on the 26 behavioral dimensions (listed earlier) which had been selected primarily to reflect differences in the extent to which the teacher fostered independence, creativity, diversity of interests, enjoyment, and intrinsic motivation among her students, ends which would be desirable among most educators of today.

Each of the behavioral dimensions was rated on a six-point scale: 3, 2, and 1 for "far," "considerably," and "slightly" above average respectively; and −1, −2, and −3 for "slightly," "considerably," and "far" below average respectively. The "average" category was omitted with the aim (by creating a forced choice condition) of avoiding the common tendency of observers (*O*s) to assign a wide variety of discriminably different behaviors to this category. Through a training program described below, an attempt was made to establish equivalent "averages" for all *O*s.

To encourage specificity of ratings, the rating scale provided space for each behavioral dimension to be rated under each of the school activities of free play, directed play, reading and/or story telling, formal instruction, outdoor activity, snack time, and cleanup activity. Owing to variation in their schedules and practices, many teachers were rated in relation to different activities, no teacher being rated for all activities. In addition to each *S* being rated on the 26 dimensions for as many of the activities as occurred during the observation period, she was also assigned an overall rating on each dimension by each *O*. An overall rating, instead of being the arithmetic average of the multiple ratings of a dimension under the different activities, represented the overall impression of an *O* for that dimension based on his interpretation of his specific ratings together with unrated behavior he may have observed or other impressions he may have gathered. In some instances the overall rating of a dimension would closely parallel the arithmetic average of ratings on that dimension and in other instances it would not. All statistical analyses were based on the overall ratings.

Training of Observers and Assessment of Interobserver Reliability

Each *O* participated in seven training sessions during which seven teachers representing Systems 1, 3, and 4 or admixtures thereof were observed and independently rated. Each observation session was followed by lengthy group discussion among the *O*s and other staff members aimed at increasing the reliability of the ratings through improving observation techniques and clarifying and standardizing meaning and usage of the rating categories.

Interjudge reliability for the 14 *O*s was assessed at two points, immediately following the last training session and immediately after completion of the experimental observations, 3 weeks later. The mean of the correlations between the ratings of every pair of judges, based both times on observations of a System 3 teacher, was .69 and .68 for the earlier and later time respectively, values which, while lower than the ideal, are as high as could be expected given the number and diversity of *O*s.

Procedure

Each *S* was observed for approximately 2½ hours by two *O*s on two occasions 1 week apart. The administrator of each Head Start program was asked to give the teachers advance notice of the dates they were to be observed; this was done in all but two instances.

The *O*s, in pairs, arrived before class began, introduced themselves, explained (with the aim of allaying the teacher's apprehension and fostering her cooperation) that the purpose of their visit was to gather examples of good teaching procedures which could be utilized as bases for subsequent training programs for Head Start teachers, and requested that they be allowed to observe but to remain as unobtrusive as possible in order to minimize the effects of their presence upon the children. To further *O*s' unobtrusiveness and simultaneously to increase the likelihood of an *S* following her planned program and, in so doing, emitting somewhat typical teaching behavior, each *S* was asked specifically not to converse with the *O*s during the class period.

The observations took place during normal classroom and playground activities on days free of special events in order to render the conditions of observations as comparable as possible for all *S*s. As a further effort toward maximizing the comparability of the judgments, each pair of *O*s was instructed to remain in the same vicinity at all times so that both would witness the same behavior of a teacher.

During the observation period *O*s independently rated *S* on all of the 26 dimensions under each of the activities that occurred during the period. Following the observation period, *O*s independently rated their overall impression of the *S* on each of the behavioral categories. Ratings were subsequently compared and discussed by each pair of *O*s and under the *single* condition where disagreement was found to exist due to different usage of the rating scale, but for no other reason, either *O* was free to make changes in his ratings. The purpose of this rather unorthodox procedure was to insure that *O*s were defining the rating categories in similar ways and in so doing to control for the variance that would otherwise result from difference in scale usage. This procedure was based on findings from two earlier studies (Harvey, 1963; Harvey & Kline, 1965) that postobservation reviews of their ratings by *O*s effectively counteracted their tendencies to drift away from the original criteria with the passage of time and the accumulation of experience. To avoid biasing the *O*s who were to observe the *S* on the second occasion, discussion concerning any *S* was restricted to the pair of *O*s who had observed her together.

Any *O* bias was further controlled and

other desired experimental controls achieved through the policies of no *O* being paired with any other *O* more than once and no *S* being observed by the same *O* more than once. Any *S* bias resulting from spread of information about the study from one teacher to another was minimized by the procedure of observing on the same day all teachers who were within the same school system.

All *O*s made their observations and ratings without any knowledge of the system classification of the teacher. Moreover, in order to prevent the *O*s from establishing particular sets and expectancies from repeated observation of *S*s from any one belief system, each *O* observed a different random order of representatives of the three belief groupings.

RESULTS

Relationships among the Rating Dimensions

The relationships among the rating dimensions were determined by a Tryon cluster analysis (Tryon & Bailey, 1965, 1966) based on the summed overall ratings of each of the 26 dimensions across the 30 teachers and the four judges. Interdimension correlations ranged from .86 to −.90, with a median correlation, without regard to sign, of .53. Fifty-five of the 325 values attained a magnitude of + or − .70 or above and all but one of these were accounted for by the interrelatedness among 15 items which cohered into one of two major clusters.

The first cluster, which may be termed *dictatorialness,* was comprised of 10 rating items. They and their factor loadings were: need for structure (.97), flexibility (−.91), rule orientation (.89), encouragement of free expression of feelings (−.86), determination of procedure (.86), use of unexplained rules (.84), punitiveness (.82), encouragement of creativity (−.78), diversity of simultaneous activities (−.77) and encouragement of individual responsibility (−.71).

The second cluster, which centered around *task orientation,* consisted of five items with factor loadings as indicated: warmth (.89), perceptiveness (.88), task effectiveness (.77), utilization of physical resources (.72), and ingenuity in improvising teaching and play materials (.65).

With the exception of the correlation between anxiety induced by judges' presence and relaxed relationships with the children, none of the correlations among the 11 dimensions not included in one of the two major factors attained the magnitude of + or − .70. Hence, while more than half of the rating items represented different ways of tapping either of two common behavior patterns, some of the items reflected fairly independent teacher behavior.

System Differences

Performance on specific dimensions. According to the hypothesis, System 4 *S*s, as classified by the TIB, should score higher than System 1 *S*s on dimensions 1–19 and lower than System 1 teachers on dimensions 20–26. A corollary hypothesis was that System 3 teachers would score between Systems 1 and 4 on all 26 dimensions.

The mean and standard deviation of the overall ratings of each system on each dimension are presented in Table 1. All values are based on the ratings of four *O*s owing to the results of a preliminary analysis of variance which showed that no significant differences existed between the ratings of the teachers at Time 1 and Time 2.

Results in Table 1 show the predicted differences between Systems 1 and 4 on all 26 dimensions. In further accord with the predictions, System 3 *S*s scored between Systems 1 and 4 on 23 of the 26 items. Page's (1963) test for ordered hypotheses for multiple treatments, a test for predictions of

TABLE 1

MEAN AND STANDARD DEVIATION FOR RATINGS OF EACH SYSTEM AS MEASURED BY THE THIS I BELIEVE TEST

Rating dimension	System 1		System 3		System 4	
	M	*SD*	*M*	*SD*	*M*	*SD*
1. Warmth	3.65	0.49	3.97	1.07	4.27	1.03
2. Perceptiveness	3.38	0.71	3.90	0.86	4.15	1.02
3. Flexibility	3.30	0.69	3.97	0.85	4.20	0.97
4. Relaxed	4.13	0.68	4.50	0.60	4.72	0.32
5. Attention to individual	3.67	0.60	4.10	0.80	4.00	0.87
6. Involvement	3.90	0.78	4.13	0.84	4.15	0.87
7. Enjoyment	3.67	0.51	4.02	0.53	4.13	0.97
8. Enlistment child participation	4.25	0.35	4.25	0.87	4.32	0.79
9. Encourage individual responsibility	3.63	0.53	4.17	0.67	4.25	0.94
10. Encourage expression of feelings	3.47	0.55	3.92	0.65	4.20	0.92
11. Encourage creativity	3.15	0.49	3.65	0.82	3.82	1.05
12. Teach new concepts	3.95	0.78	4.05	0.59	4.32	1.14
13. Ingenuity	3.30	0.64	3.85	0.88	4.27	0.73
14. Utilization of resources	3.92	0.69	4.15	0.56	4.50	0.80
15. Task effectiveness	3.82	0.44	4.22	0.48	4.35	1.06
16. Diversity of activity	3.52	0.69	4.00	0.50	4.05	1.02
17. Smoothness	3.85	0.77	4.02	0.53	4.15	0.65
18. Consistency	3.95	0.57	4.02	0.55	4.07	0.50
19. Functional explan. rules	3.38	0.66	3.85	0.74	3.77	0.72
20. Nonfunctional explan.	3.02	0.71	2.97	0.61	2.88	0.32
21. Unexplained rules	3.90	0.77	3.30	0.60	3.05	0.84
22. Rule orientation	4.47	0.76	3.70	0.87	3.20	0.91
23. Determination of procedure	4.40	0.94	4.25	0.68	3.63	0.79
24. Need for structure	4.45	1.00	3.67	0.93	3.13	1.05
25. Punitiveness	3.57	0.90	2.90	0.64	2.50	0.78
26. Anxiety	3.27	0.92	2.77	0.82	2.60	0.43

Note.—$N = 10$ for each system.

order and not of magnitude, showed that the number of dimensions on which the predicted ordering of systems was obtained was highly significant ($M = 26$, $N = 3$, $L = 361$, $p < .001$).

While no predictions were made concerning the magnitude of differences between systems on any specific dimension, a test of significance of magnitudes would yield additional relevant information. One-tailed *t* tests between systems on overall ratings indicated, on the basis of 18 degrees of freedom, that System 1 and 4 teachers differed significantly on 14 dimensions (Table 2). System 4 teachers expressed greater warmth toward the children, showed greater perceptiveness of the children's wishes and needs, were more flexible in meeting the interests and needs of the children, maintained more relaxed relationships with the children, were more encouraging of individual responsibility, gave greater encouragement to free expression of feelings, were more encouraging of creativity, displayed greater ingenuity in improvising teaching and play materials, invoked unexplained rules less frequently, were less rule oriented, were less determining of classroom and playground procedure, manifested less need for structure, were less punitive, and were less anxious about *O*s' presence.

System 3 teachers differed significantly from System 1 *S*s on eight dimensions. The former were more flexible in meeting the needs and interests of the children, were more encouraging of individual responsibility, were more task effective, permitted greater diversity of activities, invoked unexplained rules less frequently, were less rule oriented, manifested less need for structure, and were less punitive.

System 3 and System 4 *S*s differed significantly only on one dimension; the latter were significantly less determining of the classroom and playground procedures.

Dictatorialness and task orientation. The differences on the specific behavioral dimensions may be summarized and highlighted by a between-systems comparison on the two major factors extracted from the rating dimensions. With the exception of relaxed relationships with the children and anxiety from *O*s' presence, all of the specific dimensions on which Systems 1 and 4 differed significantly were contained within either the factor of dictatorialness or task orientation.

The scores on dictatorialness were 3.88, 3.32, and 3.00 for Systems 1, 3, and 4 respectively; and 3.62, 4.02, and 4.31 on task orientation. System 1 teachers were significantly more dictatorial than representatives of either System 3 ($t = 2.17$, $p < .025$) or System 4 ($t = 2.92$, $p < .005$) and at the same time significantly less task oriented than teachers from System 3 ($t = 1.73$, $p < .05$) or System 4 ($t = 2.42$, $p < .025$). There were no significant differences between Systems 3 and 4 on either factor, although System 3 teachers tended to be somewhat more dictatorial and less task oriented.

TABLE 2

t TESTS BETWEEN RATINGS OF SYSTEMS AS MEASURED BY THE THIS I BELIEVE TEST

Rating dimension	System 1–3	System 1–4	System 3–4
1. Warmth	−.87	−1.73*	−.64
2. Perceptiveness	−1.49	−1.98*	−.59
3. Flexibility	−1.95*	−2.40**	−.55
4. Relaxed	−1.31	−2.52**	−1.04
5. Attention to individual	−1.34	−.97	.27
6. Involvement	−.62	−.68	−.07
7. Enjoyment	−1.49	−1.30	−.29
8. Enlistment child participation	.00	−.27	−.20
9. Encourage individual responsibility	−2.04*	−1.83*	−.21
10. Encourage expression of feelings	−1.68	−2.15**	−.77
11. Encourage creativity	−1.66	−1.85*	−.42
12. Teach new concepts	−.32	−.86	−.68
13. Ingenuity	−1.59	−3.17***	−1.17
14. Utilization of resources	−.80	−1.72	−1.14
15. Task effectiveness	−1.94*	−1.45	−.34
16. Diversity of activity	−1.76*	−1.35	−.14
17. Smoothness	−.59	−.94	−.47
18. Consistency	−.30	−.52	−.21
19. Functional explan. rules	−1.52	−1.29	.23
20. Nonfunctional explan.	.17	.61	.46
21. Unexplained rules	1.94*	2.35**	.77
22. Rule orientation	2.12**	3.40**	1.25
23. Determination procedure	.41	2.00*	1.90*
24. Need for structure	1.80*	2.89***	1.24
25. Punitiveness	1.94*	2.86**	1.25
26. Anxiety	1.29	2.11**	.60

* $p = .05$ or $< .05$, 18 *df*, one-tailed.
** $p = < .025$, 18 *df*, one-tailed.
*** $p < .005$, 18 *df*, one-tailed.

Differences between Levels of Abstractness Derived from the CST.

The abstractness scores from the CST were divided into three levels of High, Middle, and Low, 10 scores in each group, and t tests were computed between the means of the two extreme groups for each of the 26 teacher behavior dimensions.

Results in Table 3 show that on all but four of the 26 dimensions (enlistment of child participation, teaching new concepts, smoothness of operations, and consistency of rule enforcement) the High-Abstract teachers out-performed the Low-Abstract teachers. In further support of the hypothesis, the Middle-Abstract group was rated more favorably than the

TABLE 3

MEAN AND STANDARD DEVIATION FOR RATINGS OF LOW, MIDDLE, AND HIGH ABSTRACT TEACHERS AS MEASURED BY THE CONCEPTUAL SYSTEMS TEST

Rating Dimension	Low		Middle		High	
	M	*SD*	*M*	*SD*	*M*	*SD*
1. Warmth	3.62	0.84	4.10	0.70	4.17	1.13
2. Perceptiveness	3.50	0.81	3.90	0.78	4.02	1.09
3. Flexibility	3.52	0.92	3.90	0.67	4.05	1.08
4. Relaxed	4.28	0.71	4.62	0.50	4.45	0.55
5. Attention to individual	3.78	0.56	3.98	0.88	4.02	0.87
6. Involvement	3.92	0.86	4.05	0.68	4.20	0.93
7. Enjoyment	3.78	0.65	4.00	0.50	4.05	0.94
8. Enlistment child participation	4.38	0.48	4.35	0.83	4.10	0.74
9. Encourage individual responsibility	3.98	0.64	4.00	0.76	4.08	0.94
10. Encourage expression of feelings	3.75	0.69	3.95	0.66	3.90	0.97
11. Encourage creativity	3.20	0.52	3.62	0.76	3.80	1.10
12. Teach new concepts	4.38	0.83	3.92	0.66	4.02	1.05
13. Ingenuity	3.20	0.55	3.80	0.82	4.42	0.68
14. Utilization of resources	3.88	0.72	4.42	0.43	4.28	0.85
15. Task effectiveness	4.02	0.60	4.08	0.59	4.30	0.98
16. Diversity of activity	3.50	0.66	4.05	0.64	4.02	0.95
17. Smoothness	4.15	0.76	3.80	0.63	4.08	0.55
18. Consistency	4.15	0.50	4.10	0.59	3.80	0.45
19. Functional explanation of rules	3.80	0.58	3.60	0.73	3.60	0.87
20. Nonfunctional explanation	3.15	0.76	2.95	0.35	2.78	0.46
21. Unexplained rules	3.70	0.82	3.22	0.80	3.32	0.80
22. Rule orientation	4.22	1.23	3.65	0.79	3.50	0.79
23. Determination procedure	4.40	0.86	4.12	0.92	3.75	0.73
24. Need for structure	4.15	1.19	3.70	0.98	3.40	1.10
25. Punitiveness	3.20	1.12	2.90	0.75	2.88	0.77
26. Anxiety	3.10	0.85	2.80	0.87	2.75	0.63

Note.—$N = 10$ for each group.

Low-Abstract group, but less favorably than the High-Abstract group, on 20 of the 26 dimensions. Page's (1963) test showed the frequency of predicted orderings to be significant ($M = 26, N = 3, L = 341, p < .001$).

While the number of predicted orderings was significant, the magnitude of the difference between the three CST levels, as assessed by t tests, attained significance in only five of the 78 comparisons, a frequency that departs only slightly from chance. Hence none of these t values is reported.

Abstractness scores from the CST correlated .30 ($p < .05$) with both dictatorialness and task orientation, meaning, because of the direction of scoring, that higher abstractness accompanied the tendency to be less dictatorial and more task oriented in the teaching of the Head Start children.

DISCUSSION

The results are consistent in showing that the more abstract teachers differ from the more concrete *Ss* in their teaching approaches and in the classroom atmospheres they generated for their Head Start students. Two important factors, the selective process inherent in the choice of becoming a Head Start teacher and the

1-week training program in which all teachers participated, should have operated to produce greater similarities in teaching styles and behavior toward children among our *S*s than exist among teachers selected randomly from the different belief systems and/or levels of abstractness. Despite this, the more abstract teachers in this study were clearly superior to the more concrete teachers in the extent to which they produced what are presumed to be educationally desirable atmospheres in their classrooms. Moreover, this superiority existed despite, or possibly because of, the greater teaching experience of the System 1 *S*s.

We can only conjecture at this point on the differential effect of the different atmospheres, and hence of teacher differences in concreteness-abstractness, upon the learning and behavior of the children. The answer to this, a truly significant educational question, can come only from study of the children who have been taught by teachers differing in abstractness.

Although both the TIB and CST predicted significantly differential performance in the classroom, the results point clearly to the greater superiority of the TIB. It is possible that this superiority may be reduced or eliminated by a different scoring method of the CST now being tested, one which is based on a profile of factors and categorizing of *S*s into profile groupings rather than on any single factor or summation of factors.

REFERENCES

HARVEY, O. J. Cognitive determinants of role playing. Technical Report No. 3, University of Colorado, Contract Nonr 1147(07), 1963.

HARVEY, O. J. Some cognitive determinants of influencibility. *Sociometry,* 1964, **27,** 208-221.

HARVEY, O. J. Some situational and cognitive determinants of dissonance resolution. *Journal of Personality and Social Psychology,* 1965, **1,** 349-355.

HARVEY, O. J. System structure, flexibility and creativity. In O. J. Harvey (Ed.) *Experience, structure and adaptability.* New York: Springer, 1966.

HARVEY, O. J., HUNT, D. E., & SCHRODER, H. M. *Conceptual systems and personality organization.* New York: Wiley, 1961.

HARVEY, O. J., & KLINE, J. A. Some situational and cognitive determinants of role playing: A replication and extension. Technical Report No. 15, University of Colorado, Contract Nonr 1147(07), 1965.

PAGE, E. B. Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association,* 1963, **58,** 216-230.

TIEMANN, H. A. Some social and personality determinants of reaction to sensory deprivation. Technical Report No. 14. University of Colorado, Contract Nonr 1147(07), 1965.

TRYON, R. C., & BAILEY, D. E. *The B. C. Try user's manual.* Boulder, Colorado: University of Colorado Graduate School Computing Center, 1965.

WARE, R., & HARVEY, O. J. A cognitive determinant of impression formation. *Journal of Personality and Social Psychology,* in press.

WHITE, B. J., & HARVEY, O. J. Effects of personality and own stand on judgment and production of statements about a central issue. *Journal of Experimental and Social Psychology,* 1965, **1,** 334-347.

(Received March 8, 1966)

Journal of Educational Psychology
1966, Vol. 57, No. 6, 382–388

COGNITIVE TRANSFER IN VERBAL LEARNING[1]

JAMES H. REYNOLDS
Colgate University

In a 2-stage transfer paradigm, a cognitive-learning group ($N = 10$) was first familiarized with a map drawing which presented verbal material embedded in an organized perceptual structure, and then learned 8 simple sentences giving factual material related to the previously-studied map. Control condition groups ($N = 10$ each) learned the same sentences either after exposure to various components of the total map structure or without prior exposure to any material. The cognitive group demonstrated superior learning ($p < .05$) of the sentences relative to each of the control groups, but no differences in learning were found among the various controls. The results were interpreted as indicating (a) that the integration of the verbal and perceptual stimuli into a total structure, rather than these components separately, was responsible for the transfer observed, and (b) that cognitive organization may facilitate simple verbal learning as well as complex and conceptual learning.

Tolman (1948) has attributed much of the learning of animals and men to their reliance upon a "cognitive map," into which new learning is meaningfully assimilated. His figurative expression refers to an assumed mental organization or structure, built up from past experience, which makes the learner's surroundings meaningful. It is hypothesized that if new situations, requiring learning, can be incorporated into a meaningful structure, learning is relatively easy; if not, learning is by rote, and more difficult. This basic hypothesis has been elaborated by other theorists (e.g., Ausubel, 1963; Bruner, 1957), and forms a theoretical basis for such school-learning practices as understanding versus drill, the use of visual aids to introduce new topics, and other methods designed to broaden the student's conceptual overview of academic subjects.

Experimental attempts to demonstrate the effect of cognitive structure upon human learning ordinarily employ a two-stage transfer paradigm. In the first stage the learner is familiarized with materials which provide opportunity for the formation of a meaningful organization or structure, and in the second stage new material which bears some relationship to that of the first stage is presented for learning. Observation of positive transfer in the second-stage learning task is taken as evidence that a cognitive structure was established by the first-stage treatment.

Ausubel (1963, Ch. 5) has reviewed the extensive research literature demonstrating by this method the effects of cognitive structure upon human learning. Most of this research, however, has employed strictly verbal materials in the first stage (e.g., essays, rule-learning), and tested for transfer to such broad and complex tasks as the learning of conceptual academic subject matter or the solution of difficult problems and puzzles. The effects of perceptual organization, and cognitive effects upon the learning of simple verbal tasks, have received little study. Also, in many investigations the cognitive condition was compared only with control conditions which in the first stage received no treatment relevant in any way to the second-stage learning task. Since cognitive material is usually complex in structure and consists of a variety of interrelated stimulus components, it is possible that the positive transfer obtained may be attributable not to the organization or meaningfulness of the first-stage task per se, but rather to certain discrete elements of that task. Were this the case, findings of positive transfer might be explained better by assuming a discrete stimulus-response (S-R) association process, rather than inferring a cognitive mechanism. Unless appropriate control conditions are employed, it is difficult to make a clear decision between these alternative explanations of the positive transfer obtained.

---

[1] This research was supported by a research grant from the United States Office of Education, Contract No. OE-6-10-43.

The present study was directed at extending and clarifying three aspects of the cognitive research available to date. First, the cognitive material presented in the first stage of the transfer paradigm was an integrated perceptual-verbal structure rather than the organized verbal structure often employed. The success of visual aids in facilitating learning, and the importance attributed to visual stimuli by Gestalt and other cognitive theories, suggest that a visual organization should have strong positive transfer effects to the learning of related verbal material. Taking Tolman quite literally, the cognitive material employed was a map-like drawing on which verbal stimuli were printed. Second, the learning task following the cognitive presentation consisted of learning specific facts related to the previously shown map, rather than learning broad concepts. This type of task was employed for two reasons: (*a*) Much of academic learning can be characterized as the assimilation of new factual material into an existing cognitive structure (Ausubel, 1963), and (*b*) the simplicity of the task made it possible to analyze transfer effects in terms of S-R, as well as cognitive, relationships.

Finally, an attempt was made to determine the locus of the expected transfer by employing five control conditions. The map used in the cognitive treatment integrated several discrete stimulus components—for example, the drawings of certain objects in a specific relationship to each other, various verbal stimuli, and positional relationships between the verbal stimuli and the pictured objects. It was considered that only certain of these several stimulus aspects, and not the total structure itself, may be responsible for positive transfer. Also considered was the possibility that the map structure might serve simply to establish implicit verbal S-R associations which could mediate later learning and result in positive transfer. To evaluate the possible effects of these separate components upon second-stage learning, various control groups received in the first stage either the map and the verbal stimuli separately (i.e., not integrated into a meaningful structure), the verbal stimuli alone in varying configurations, verbal stimuli which provided only discrete S-R mediating associations, and the usual transfer control of no treatment prior to second-stage learning.

METHOD

Subjects

The subjects (*Ss*) were 60 college men at Colgate University who participated to fulfill a requirement for an introductory psychology course. All *Ss* were assigned to experimental conditions according to a predetermined random order as they appeared individually for the experiment.

Design and Procedure

Stage II task. A total of six groups were employed ($N = 10$ per group), *Ss* in each group being exposed to two experimental stages. The Stage II task was to learn the following eight simple sentences, each beginning with the three-letter initials of a hypothetical person and ending with a statement of that person's occupation: KOT is a pilot, NEB is a gas station attendant, BAF is a shopkeeper, PUM is a truckdriver, DOS is a policeman, RAZ is a cook, FER is a brakeman, and TUK is a farmer. The initials used were CVC's of 85–100% association value taken from Glaze's (1928) list. These sentences constituted factual material about eight hypothetical persons, and were also considered as eight paired associates, each with the initials serving as the S-term and a word of a certain class (occupations) serving as the R-term.

Following various Stage I treatments (described below), *S*s in all groups were seated before a Stowe memory drum set in a 3 × 4-foot black screen and instructed to learn the eight sentences as quickly as possible. Seven learning trials were then administered. On each trial, the sentences were first presented, one at a time, at a 2-second rate in the drum window for silent study. After this study period the drum was stopped and *S* was given 75 seconds to write on a test sheet as many of the sentences as he could recall. Then the test sheet was returned and 15 seconds later a new trial began, again with a silent study period via the memory drum followed by a test. The sentences were presented in randomly varied orders on each trial to prevent serial learning, and *S* was told he could write the sentences in any order he wished on each test. The relatively long test period provided ample time for even the slowest writer to write all the sentences he could recall on every trial.

Stage I treatments—Cognitive Structure Group. The purpose of the Stage I treatment administered to the Cognitive Structure Group was to impose experimentally a meaningful and integrated cognitive representation of stimulus components related to the Stage II learning task. The integrative representation used was an 8½ × 11-inch map sketch depicting a highway intersection and its surroundings, including an airstrip, a gas station, a shopping center, a diner, a farm, and a railroad. A tractor-trailer and a police patrol car were drawn on the highways. Beside each of these eight major components of the map structure was typed a set of three-letter initials, the stimuli for the Stage II task. Figure 1 illustrates the map and the initials embedded in their appropriate positions.

The *S* was instructed to study the initials on this map for 30 seconds and memorize them so that he could reproduce them in their proper positions on a later test. After 30 seconds the map was removed and *S* had 1 minute to recall and write on a test map the missing initials in their correct places. Following the first test two more 30-second study periods were given, with a 1-minute test between each, providing adequate time for familiarization with the initials and their correct placement. At the end of this Stage I treatment, it was presumed (*a*) that *S* would be familiar with all of the initials to which he would later (Stage II) have to associate occupations, and (*b*) that the initials would be embedded in a representative and meaningful cognitive structure (i.e., the integrated map) which would aid in Stage II learning.

Map-List Group. The Cognitive Group received in Stage I familiarization of both verbal and perceptual stimuli in a specific relationship. The remaining groups were essentially control conditions, providing at Stage I varying degrees of familiarization with either verbal stimuli alone, nonintegrated verbal and perceptual stimuli, or neither type of stimuli, but never all of the components of the total map stimulus in an integrated relationship.

Stage I for the Map-List Group consisted of presenting to *S* for study (*a*) the map shown in Figure 1 *without* any embedded initials, and (*b*) a list of the eight sets of initials on a separate 3 × 8-inch strip of paper, typed one beneath the other with double spacing between each. Instructions were to study the map and the initials for three 30-second periods; with a 1-minute test after each study period, in which *S* was to write down all the initials he could recall, in the correct list order, on a 3 × 8-inch test strip which contained only eight lines on which to write the correct responses. The separate map remained in view during both the study and test periods, but at no time were any initials placed on any part of the map. Thus, *S* was familiarized with the initials and with the map, but these two components were never meaningfully related in a single structure.

List Group. The Stage I treatment for the List Group was identical to the Map-List Group, except that the separate map was omitted from both the study and the test periods. Thus this treatment provided familiarization with the discrete verbal stimuli, but these stimuli were neither embedded

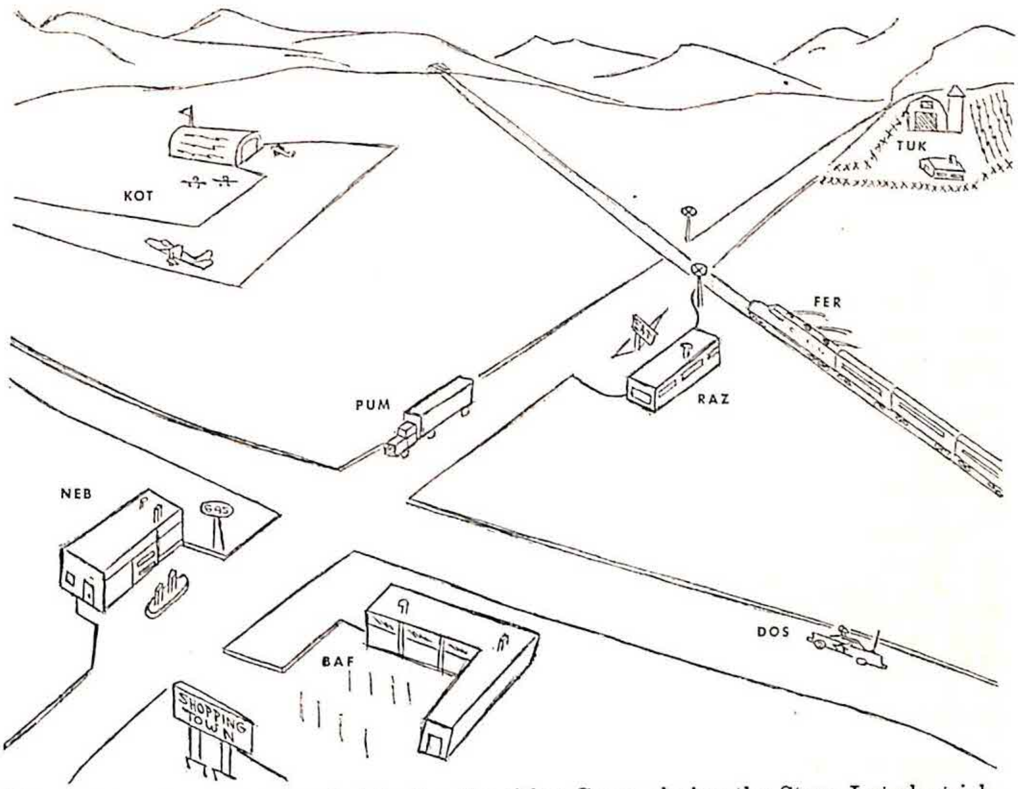


FIG. 1. Map structure presented to the Cognitive Group during the Stage I study trials.

in a perceptual structure nor was a potentially useful structure presented.

Positioned-List Group. Asch, Hay, and Diamond (1960) have shown that serial learning may be affected by the physical configuration or positioning of the items being learned. Therefore it was considered possible that the specific positioning per se of the initials on the map (Figure 1), for example, KOT in the upper left corner, DOS in the lower right corner, etc., might isolate the stimuli in a manner which would facilitate learning in Stage II regardless of whether the map was presented or not. As a control to evaluate this possibility, the Position-List Group received for study in Stage I an 8½ × 11-inch sheet of paper on which the eight initials were printed in the same relative positions as shown on the map in Figure 1, but with the map omitted. This group also received three 30-second study periods, interspersed with three test periods in which *S* wrote all the initials he could recall in their appropriate positions on an 8½ × 11-inch test sheet containing only lines on which to write the correct responses. This Position-List Group was similar, then, to the List Group, in that it received familiarization of the initials in the absence of the map structure; but differed from the List Group in that it learned the initials in a specific configuration identical to the map positions, rather than in a vertical list.

S-R Group. One alternative interpretation of the function of the map in Figure 1 is that it provides not a "cognitive structure" but (*a*) familiarization with the verbal stimuli, which may transfer positively to the Stage II task (Goss & Nodine, 1965, Ch. 6), and (*b*) an opportunity to learn a series of discrete verbal S-R associations, which may mediate Stage II learning and thus produce positive transfer. An example of the latter interpretation is that study of the upper left corner of the map may yield the implicit verbal association "airstrip–KOT" which, at Stage II, will mediate the association "KOT–pilot" because of the extraexperimental association existing between "airstrip" and "pilot." In transfer paradigm form, this is an A-B, B-A′ relationship; and Houston (1964) has reviewed evidence suggesting that this paradigm should yield positive transfer.

To evaluate the adequacy of this alternative explanation of the map function, an

S-R Group received in Stage I a blank 8½ × 11-inch study sheet on which was typed, in the same relative positions as on the map in Figure 1, the following eight verbal paired-associates: airstrip–KOT, gas–NEB, shopping center–BAF, tractor-trailer–PUM, patrol car–DOS, diner–RAZ, train–FER, farm–TUK. The *Ss* in this group also received three 30-second study periods, interspersed with 1-minute tests during which they wrote the initials in their correct positions on a test sheet which contained the eight nouns alone, typed in their proper places on the sheet. Thus the S-R Group received both familiarization with the initials and an opportunity to form discrete verbal associations of the type presumably available to the Cognitive Group, but no potentially integrative relationship among the items was provided.

Rest Group. A final control group learned the list of sentences in Stage II, but received no Stage I treatment at all. This Rest Group was used to evaluate the effects of all other Stage I treatments upon Stage II learning.

RESULTS

Stage I learning. Despite the ease of the Stage I tasks, one to three *Ss* in every group failed to learn all of the initials in the three study periods administered. The mean number of correct initials given on the last test in Stage I ranged from 7.00 to 7.60 for the five groups receiving the Stage I treatment. A preliminary analysis of variance verified that the differences among these means were not significant, $F(4,45) < 1.00$, $p > .10$; indicating that while Stage I learning was not perfect, the conditions receiving the various Stage I treatments did not differ in degree of familiarization with the initials at the time Stage II began.

Stage II learning. The means and *SD*'s for the total number of correct responses made by each group over the seven trials of Stage II learning are presented in Table 1. The overall difference among the means, tested by analysis of variance, was highly significant, $F(5,54) = 3.77$, $p < .01$. Comparisons of individual control-group means, using the Newman-Keuls method, indicated no significant differences among the five groups which did not receive the Cognitive Group treatment. Neither familiarization of the initials alone, familiarization of the initials in the presence of the separate map structure, nor the provision for verbal mediating associations, resulted in positive transfer when compared to the Rest Group receiving no Stage I treatment at all. Turning to the Cognitive treatment, however, further Newman-Kuels comparisons showed that the difference between the Cognitive Group and each of the control conditions was significant beyond the .05 level of probability. Thus positive transfer was obtained, but only when the various components of the Stage I task were integrated into an organized perceptual-verbal structure.

Figure 2 presents the learning curves for each group over the seven Stage II learning trials, illustrating the consistent superiority of the Cognitive

TABLE 1

MEANS AND SD'S FOR TOTAL NUMBER OF CORRECT RESPONSES OVER SEVEN STAGE II TRIALS

	Groups					
	Cognitive	Map-list	List	Position-list	S-R	Rest
M	49.5	39.8	39.6	33.9	41.6	38.2
SD	7.25	8.72	6.90	8.10	11.40	7.02

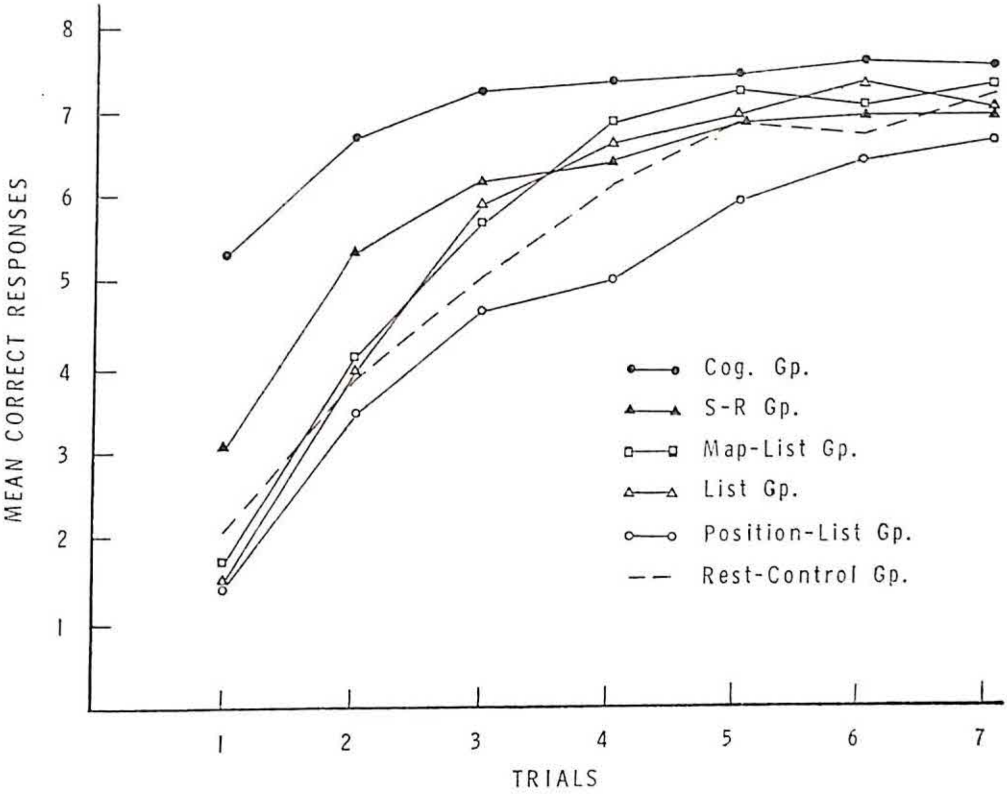


FIG. 2. Stage II learning curves for all groups.

Group beginning with the first trial. Individual t tests indicated that the Trial 1 mean for the Cognitive Group was significantly higher than the second highest S-R Group mean, $t(18) = 2.59$, $p < .02$. The S-R Group, although it showed an initial trend toward higher performance than the other controls, did not differ significantly from the Rest Group on the first trial, $t(18) = 1.32$, $p > .10$. Differences among the remaining four control groups were also nonsignificant. Following the first trial the groups tend to converge, the Cognitive Group approaching the test ceiling by Trial 3 and the others approaching ceiling at a later point. Although learning was rapid in all groups, none of the conditions matched or exceeded the Cognitive Group after seven trials of learning.

DISCUSSION

The data clearly indicate that prior exposure to an integrated and potentially meaningful perceptual structure may provide positive transfer to a related task in which simple factual materials are to be learned. The evidence is consistent with past findings demonstrating positive transfer from meaningful verbal structures to more complex tasks such as problem solving and the learning of conceptual materials, and confirms their results with another type of structure and a simple but common learning situation.

The S-R Group did not differ significantly from the other controls, despite a trend toward superior performance on the initial trials of the Stage II task. Thus the A-B, B-A′ paradigm, when used with discrete verbal associative materials, was found to be ineffective in producing positive transfer. In contrast the Cognitive Group, which received the equivalent A-B, B-A′ sequence but with integrated perceptual stimuli in the A-B stage, demonstrated strong positive transfer. Whether the superiority of the latter group was due to the perceptual aspect or the integrative aspect of the Stage I stimuli can not be determined from this comparison, but regardless of the specific locus there is a strong suggestion that the integrated perceptual stimuli presented to the Cognitive Group provided more than a series of discrete verbal mediators.

The Cognitive Group was also consistently superior to the three control groups which received various components of the total map structure during Stage I; but none of these three controls demonstrated positive transfer relative to the Rest Group, which received no Stage I treatment at all. This lack of positive transfer in the component control groups suggests that the transfer observed for the Cognitive Group was not attributable to the specific map components presented, either alone or in combination, but rather to the organization of the various components into a single structure. It is this integration or interdependence of the discrete components of an experience which is ordinarily implied by the term "meaning" as it is used in describing cognitive structure. The data indicate that it was the organizational or "meaning" aspect of the Stage I task, rather than the familiarization with either the initials alone or the initials and the map separately, which contributed to the positive transfer demonstrated by the Cognitive Group. The fact that this transfer was obtained with simple materials which were easily learned by rote makes it reasonable to conclude that, for human learners at least, this organizational aspect is a powerful factor in even basic learning processes.

REFERENCES

ASCH, S. E., HAY, J., & DIAMOND, R. M. Perceptual organization in serial rote learning. *American Journal of Psychology,* 1960, **73,** 177–198.

AUSUBEL, D. P. *The psychology of meaningful learning.* New York: Grune & Stratton, 1963.

BRUNER, J. S. Going beyond the information given. In *Contemporary approaches to cognition: Report of a symposium at the University of Colorado, May 12–14, 1955.* Cambridge: Harvard University Press, 1957.

GLAZE, J. The association value of nonsense syllables. *Journal of Genetic Psychology,* 1928, **35,** 255–269.

GOSS, A. E., & NODINE, C. F. *Paired-associates learning.* New York: Academic Press, 1965.

HOUSTON, J. P. Verbal transfer and interlist similarities. *Psychological Review,* 1964, **71,** 412–414.

TOLMAN, E. C. Cognitive maps in rats and man. *Psychological Review,* 1948, **55,** 189–208.

(Received March 17, 1966)

Guidance counselling is basically a matter of people.

But it helps to have a well-stocked library of scientifically constructed guidance materials that can assist you with your special needs.

SRA is the acknowledged leader in the commercial publishing of school guidance materials, with over 20 guidance programs or series and numerous guidance products for all elementary and secondary levels.

SRA offers guidance materials in the areas of:

Information and Development	**personal**	**A Book About Me (K-1)**
	educational	**Learning to Think Series (K-1)**
	vocational	**Charting Your Job Future (9-12)**
	occupational	**Career Information Kit (10-12)**
Reading		**Lift-Off to Reading (1-6)**
Social Studies		**Our Working World (1-3)**
Guidance Books and Pamphlets	**professional**	**Guidance Activities for Secondary School Teachers**
	student	**Junior Guidance Series (7-9)**

We will be pleased to send you information about the full line of SRA guidance materials.

SRA

Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A Subsidiary of IBM

Harper & Row

Learning and Human Abilities:

Educational Psychology, Second Edition

Herbert J. Klausmeier and William Goodwin

Almost completely rewritten and thoroughly revised, this Second Edition continues to emphasize the concept of emerging human abilities, thus integrating the treatment of growth and learning. Chapters 1 and 2 now include new material on learning theories; Chapters 7–13 present Instructional Models that represent a streamlining of the Principles of Instruction. Preceding each section is an advance organizer providing an overview of material to follow. Available to teachers: *Student Evaluation Guide* *720 pages, $8.95*

Student Workbook

William Goodwin, Robert Conry, and Herbert J. Klausmeier

This supplement consists of a student workbook, comprised of questions that encourage high-level intellectual abilities of students. Includes an adjunct program providing forty multiple-choice questions. *$3.25*

Readings in Learning and Human Abilities:

Educational Psychology

Richard E. Ripple

Forty-nine unabridged readings from twenty-three journals, for use with Klausmeier and Goodwin's *Learning and Human Abilities: Educational Psychology* or with other introductory educational psychology texts. Although emphasis is on articles reporting empirical research studies, some theoretical articles and survey articles reviewing research are included. A balance is achieved as among articles dealing with the elementary, secondary, and college levels. A specially prepared article will aid students in handling the statistics in the selections. Introduction to each of the eight sections; interpretive headnote for each reading.

586 pages, paper, $4.95

Harper & Row, Publishers
49 East 33d Street, New York 10016

New and Forthcoming from McGraw-Hill

CONDITIONING AND INSTRUMENTAL LEARNING

By *Wendell I. Smith* and *J. William Moore*; both of Bucknell University. 192 pages; $4.95 (clothbound), $2.50 (soft-cover). The first in a new series of programmed psychology books which provide extensive coverage of key topics taught in first-year psychology. Units can be assigned in lieu of text readings to reduce the need for classroom discussion.

ADOLESCENT DEVELOPMENT, Third Edition

By *Elizabeth B. Hurlock*, The Graduate School of Education, University of Pennsylvania. *McGraw-Hill Series in Psychology*. 752 pages; $8.95. In an extraordinarily thorough and comprehensive treatment, Dr. Hurlock covers the traditional topics and extends the material to emphasize changing cultural patterns and the emergence of new problems for today's adolescent. An instructor's manual is available.

THE PSYCHOLOGY OF LEARNING, Third Edition

By *James Deese* and *Stewart H. Hulse*; both of The Johns Hopkins University. *McGraw-Hill Series in Psychology. Available Winter.* Provides as representative a picture as possible of the contemporary concern with the psychology of learning. Broad coverage and detailed exposition of basic problems are the most important features of the book.

SCALES FOR THE MEASUREMENT OF ATTITUDES

By *Marvin E. Shaw* and *Jack M. Wright*; both of the University of Florida. *McGraw-Hill Series in Psychology. Available Winter.* Here are all the worthwhile attitude scales in a single source. An extensive bibliography is included.

SHORT-TERM PSYCHOTHERAPY AND STRUCTURED BEHAVIOR CHANGE

By *E. Lakin Phillips*, George Washington University; and *Daniel N. Wiener*, Veterans Administration, St. Paul and University of Minnesota. 256 pages; $7.95. This *systematic* approach to shortening therapy is accomplished by showing the fundamental importance of *behavioral* techniques.

EXPERIMENTAL METHODS AND INSTRUMENTATION IN PSYCHOLOGY

By *Joseph B. Sidowski*, San Diego State College. *McGraw-Hill Series in Psychology.* 768 pages; $15.00. This advanced-level contributed book describes some of the experimental methods and instrumentation used in psychology. No other single source in the field contains this information.

SENSITIVITY TO PEOPLE

By *Henry Clay Smith*, Michigan State University. 224 pages; $6.95. Shows how sensitivity to people can be improved. Analyzes the effectiveness of present training, ways of measuring sensitivity, qualities of sensitive people, development of sensitivity, and ways of training people to eliminate the errors they make in judging others.

PSYCHOMETRIC THEORY

By *Jum C. Nunnally*, Professor of Psychology, Vanderbilt University. *McGraw-Hill Series in Psychology. Available Spring.* Provides comprehensive coverage of the major issues in psychological measurement. Intended for graduate courses.

Reserve your examination copies today.

McGRAW-HILL BOOK COMPANY

330 West 42nd Street, New York, New York 10036

AVAILABLE IN JANUARY

Essentials of Learning
Second Edition

By Robert M. W. Travers, Western Michigan University

This thoroughly updated edition of a highly respected text provides unusually broad coverage of the field of learning and discusses individual topics within the field in detail. Thus, not one theory of learning is presented, but many, each firmly rooted in research, providing the student with significant and current data on every aspect of learning. The author analyzes the possible implications that modern studies of learning may have for curriculum planning, for teaching methods, and for general educational planning. An extensive bibliography of over 550 references and an index of both names and subjects are included at the end.

Special Features:

*All discussion is fully documented and based on current research

*The physiology of information transmission, cognitive learning, and perception are all discussed.

*Chapters on verbal learning, motivation, and the nervous system have been updated and strengthened.

1967, approx. 544 pages, $7.95

Write to the faculty service desk for examination copies

THE MACMILLAN COMPANY

866 Third Avenue, New York 10022

The Psychological Foundations of Education Series

Edited by Victor H. Noll, Michigan State University

This paper bound series provides a deep examination of specific areas in educational psychology that are not given thorough treatment in many comprehensive textbooks now in use. The series offers a convenient source of up-to-date and expertly written supplementary material. Used in conjunction with a comprehensive text, these books will increase the impact and effectiveness of current courses in the psychological foundations sequence. Theoretical material is an important part of each book, but emphasis is centered on practical applications.

Psychology of Adolescence for Teachers
By Glenn Myers Blair, and R. Stewart Jones; both of University of Illinois
1964, 128 pages, $1.50

Psychology of the Child in the Classroom
By Don C. Charles, Iowa State University
1964, 96 pages, $1.25

The Psychology of Learning in the Classroom
By Robert C. Craig, Marquette University
1966, 85 pages, $1.25

The Mentally Retarded Child in the Classroom
By Marion J. Erickson, Ypsilanti Public Schools
1965, 128 pages, $1.50

Problem Solving in the Classroom
By Bryce B. Hudgins, Washington University
1966, 96 pages, $1.25

Teacher Self-Evaluation
By Ray H. Simpson, University of Illinois
1966, 96 pages, $1.25

Guidance in the Classroom
By Ruth Strang, University of Arizona; and Glyn Morris, Assistant Supervisor of Guidance and Curriculum, Lewis County, New York
1964, 128 pages, $1.50

Gifted Children in the Classroom
By E. Paul Torrance, University of Minnesota
1965, 112 pages, $1.25

Write to the faculty service desk for examination copies.

THE MACMILLAN COMPANY 866 Third Avenue, New York 10022

Reading In High Gear helps him learn how to read and how to live.

READING IN HIGH GEAR is a contemporary learning program tailored to meet the demands of the more mature semi-literate and illiterate learners. Without the ability to read, these people live and work at a disadvantage. For them, learning to read is the first step on the road to a better life. The stories in READING IN HIGH GEAR aim at developing healthy social attitudes and acceptable, realistic ways of adjusting to society. Story situations are designed to be both interesting and relevant to the experience of the participant.

The program was originally developed by Dr. Myron Woolman, primarily for the culturally disadvantaged adolescent and adult. Its method is based on the progressive choice learning theory which relies on single responses and dissimilarity of appearance in material to facilitate learning. Assuming only that the participant can speak English, the program helps him respond accurately to the printed form of words that are already meaningful to him in their spoken form.

READING IN HIGH GEAR supplies teachers with a wide range of reading materials suited to the individual needs of remedial and pre-dropout students.

Give young and mature adults a chance to catch up with yesterday. Write for a detailed brochure, or let an SRA Staff Associate tell you more about READING IN HIGH GEAR.

SRA

Science Research Associates, Inc.
259 E. Erie St., Chicago, Ill. 60611

A Subsidiary of IBM

LABORATORY AIDS *in* TEACHING PSYCHOLOGY

For Classical Conditioning

"LABORATORY STUDIES of the CONDITIONED RESPONSE" by Howard S. Hoffman, Associate Professor of Psychology, Pennsylvania State Univ. A student manual and workbook to accompany the ANALOGUE LAB. Newly published.

The Hoffman ANALOGUE LAB

- A book-sized electrical analogue of a Pavlovian laboratory. Self contained.
- Permits the student to focus on the problems of design, analysis, and reporting of experiments rather than on handling apparatus and animals.
- Complex enough to simulate extinction reminiscence, spontaneous recovery, secondary reinforcement through higher ordered conditioning.
- Inexpensive enough for every student of introductory psychology to use it 2 hours a day.

For Operant Conditioning:

STUDENT LAB SYSTEMS

- With modular test chamber (for rat or pigeon)
- With automated stimulus presentation & response-sensing.
- With automated counting, timing, and scheduling of reinforcement.
- Uses research-grade equipment, expansible to research level lab work. Available from stock.
- Scheduling system with relays or solid state.

FORINGER
& CO, INC.

BEHAVIORAL RESEARCH SYSTEMS ROCKVILLE MARYLAND 20850

AERA

Published Four Times a Year....

January, March, May, November

AMERICAN EDUCATIONAL RESEARCH JOURNAL

Official Publication of the American Educational Research Association

- Original reports of experimental and theoretical studies in education and occasional reviews and tests.
- Annual subscription, $6.00.
- Address manuscripts and editorial correspondence to the Editor, **Robert W. Heath**, *American Educational Research Journal*, P.O. Box 6566, Stanford, California 94305.
- Mail subscriptions to **American Educational Research Association**, 1201 Sixteenth Street, N.W., Washington, D.C. 20036.

Special Savings for AMERICAN PSYCHOLOGICAL ASSOCIATION members. . . . *American Educational Research Journal*, 1964 and 1965, Volumes I and II; each volume of four issues, $3.50. For individual copies of each or both, send the coupon with check or money order for the full amount, payable to the National Education Association.

Publications Sales Section, National Education Association

1201 Sixteenth Street, N.W. Washington, D. C. 20036

Payment of ______ ($3.50 for *each* vol.) is enclosed for Vol. I, Stock #048-01714, and Vol. II, Stock #048-01716 of the *American Educational Research Journal.*

Name. . . . .

Address . . . .

City . . . . State. . . . . Zip. . . . . . .

the
journal
of
special
education

partial contents, vol. 1, no. 1

...drug effects on learning in children... new directions for understanding school failure... cigarette smoking among junior high school students... social reinforcement system for the classroom management of behavior disorders... the "teacher-mom" program evaluated... the questionable role of specialists in special education programs... the devereux elementary school behavior rating scales... special education teaching in the occupational structure...

Please send manuscripts to Lester Mann, at the address below. A limited number of complimentary copies of the inaugural issue are available—or you may subscribe using the coupon below.

the **journal** of **special education**
Box 455, Ft. Washington, Penna., 19034

name.........................
address.....................
city.......... state....... zip....
....please send me complimentary copy

(student $8, individual $9, institution $10)

Oxford
University Press

Fundamentals of Measurement:

Techniques and Practices
Second Edition

By **N. M. Downie**, Purdue University

Ideally suited to the one-semester course, the Second Edition of this well-known text provides a concise and practical introduction to the techniques and practices of measurement. Major emphasis is placed on those principles of measurement which are fundamental to all areas of human learning with attention focused on the most representative tests in the major fields.

The text is organized in three main parts. Part I includes chapters which explain basic statistical principles that govern measurement and evaluation, reliability and validity, and general procedures used in the selection, administration, and scoring of standard tests. Part II contains chapters on achievements tests, including materials on item writing; the essay examination and item analysis techniques. Part III includes chapters on the use and interpretation of measurement and appraisal techniques applied to the evaluation of intelligence, special abilities, personality, interest, attitudes, and socio-economic status. In addition to numerous tables and charts, the text has a bibliography, a complete index, and a glossary of technical terms.

January 1967 484 pp. prob. $6.50

Oxford University Press
200 Madison Avenue
New York, New York 10016

Emotional Disturbance and School Learning: a Book of Readings

Donald H. Clark, Hunter College
of New York City University
Gerald S. Lesser, Harvard University

"School learning is not simply learning the three R's; it is learning to master many things as diverse as arithmetic, hostility, spelling, and tenderness."

This compact anthology of 26 articles introduces the reader to current research on the problem of school learning. It stresses the value of a critical approach toward research and the scientific method and emphasizes that research provides information, not answers. The reader is encouraged to sift and weigh the various psychological approaches in the articles in order to develop his own critical faculties.

EMOTIONAL DISTURBANCE AND SCHOOL LEARNING affords excellent supplementary reading for courses in child development, educational psychology, mental hygiene, counselling and guidance, personality, and methods in education. For the school guidance counsellor it can help clarify the confusion surrounding the "scientific method" and, with its ample bibliography, serve as a source book for further investigation of emotional disturbance in the child. 294 pp., paperbound, $3.25 list, $2.60 net.

Write for your free examination copy of EMOTIONAL DISTURBANCE AND SCHOOL LEARNING.

SRA

Science Research Associates, Inc.
College Department
259 E. Erie St., Chicago, Ill. 60611

A Subsidiary of IBM